10-27-00

A

**Patent Application**
**Docket No.** 48279-3USPT

## IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re patent application of:

Lee A. Bulla, Jr. and Mehmet Candas

For: GENE MINING SYSTEM AND METHOD

BOX APPLICATION
Assistant Commissioner
for Patents
Washington, D.C. 20231

CERTIFICATE OF MAILING BY EXPRESS MAIL

"EXPRESS MAIL" Mailing Label No. EL524961412US
Date of Deposit: October 25, 2000
*I hereby certify that this paper or fee is being deposited with the U.S. Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231*

Type or Print Name: Carol Marstaller
Carol Marstaller
Signature

Sir:

## PATENT APPLICATION TRANSMITTAL LETTER

Transmitted herewith for filing, please find the following:

X Specification, claims and abstract of the above-referenced patent application (total of 83 pages)

X 14 sheet(s) of drawing(s) (__formal/ x informal).

X Combined Declaration and Power of Attorney (signed)

_ An Assignment of the invention to: ________________________

X A verified statement claiming small entity status under 37 CFR 1.9(f) and 1.27(b) - Independent Inventor.

X Priority is claimed under 35 U.S.C. § 119 based on filing in U.S. Patent and Trademark Office .

| | Application No. | Filing Date |
|---|---|---|
| (1) | 60/161,527 | 10/26/99 |
| (2) | 60/161,571 | 10/26/99 |
| (3) | ____ | ____ |

__ (No.) Certified copy (copies) __ are attached; or ____ were previously filed on ____.

The filing fee has been calculated as shown below:

| FOR: SMALL ENTITY | NO. FILED | NO. EXTRA | RATE | FEE |
|---|---|---|---|---|
| BASIC FEE | | | | $355 |
| TOTAL CLAIMS | 82 - 20 | 62 | $9 | $558 |
| INDEPENDENT CLAIMS | 8- 3 | 5 | $40 | $200 |
| MULTIPLE DEPENDENT CLAIM(S) PRESENTED | | | $135 | $135 |
| TOTAL FEE: | | | | $1,248 |

X Please charge my Deposit Account No. 10-0447 in the amount of $1,248.00. This sheet is attached in duplicate.

___ A check in the amount of $____________ is attached. Please charge any deficiency or credit any overpayment to Deposit Account No. 10-0447.

X The Commissioner is hereby authorized to charge payment of the following fees associated with this communication or credit any overpayment to Deposit Account No. 10-0447 This sheet is attached in triplicate.

X Any additional filing fees required under 37 CFR 1.16 including fees for presentation of extra claims.

X Any additional patent application processing fees under 37 CFR 1.17 and under 37 CFR 1.20(d).

X The Commissioner is hereby authorized to charge payment of the following fees during the pendency of this application or credit any overpayment to Deposit Account No. 10-0447. This sheet is attached in duplicate.

X Any patent application processing fees under 37 CFR 1.17 and under 37 CFR 1.20(d).

_ The issue fee set in 37 CFR 1.18 at or before mailing of the Notice of Allowance, pursuant to 37 CFR 1.311(b).

X Any filing fees under 37 CFR 1.16 including fees for presentation of extra claims.

JENKENS & GILCHRIST

Date: October 25, 2000

Raymond Van Dyke
Registration No. 34,746

JENKENS & GILCHRIST
1445 Ross Avenue, Suite 3200
Dallas, Texas 75202-2799
214/855-4708

09699110250

DOCKET NO. 48279.3

**VERIFIED STATEMENT CLAIMING SMALL ENTITY STATUS**
**(37 CFR 1.9(f) & 1.27(b)) - INDEPENDENT INVENTOR**

Applicant or Patentee: Bulla, Jr. et al.

Serial No. Unknown

Filed: Unknown

Title: GENE MINING SYSTEM AND METHOD

As a below named inventor, I hereby declare that I qualify as an independent inventor as defined in 37 CFR 1.9(c) for the purposes of paying reduced fees to the Patent and Trademark Office with regard to the invention as described in:

[X ] the specification filed herewith.

[ ] application serial no. ________________, filed________________.

[ ] patent no. ______, issued ___________.

I have not assigned, granted, conveyed or licensed and am under no obligation under contract or law to assign, grant, convey or license, any rights in the invention to any person who could not be classified as an independent inventor under 37 CFR 1.9(c) if that person had made the invention, or to any concern which would not qualify as a small business concern under 37 CFR 1.9(c) or a nonprofit organization under 37 CFR 1.9(c).

Each person, concern or organization to which I have assigned, granted, conveyed, or licensed or am under an obligation under contract or law to assign, grant, convey, or license any rights in the invention is listed below:

[X] no such person, concern, or organization

[ ] persons, concerns or organizations listed below*

* NOTE: Separate verified statements are required from each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b))

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

Name of Inventor:
LEE A. BULLA, JR.

Date: 25Oct00

Signature of Inventor

Name of Inventor:
MEHMET CANDAS

Date: 10/25/2000

Signature of Inventor

Patent Application
Docket No. 48279-3USPT

CERTIFICATE OF MAILING BY EXPRESS MAIL

"EXPRESS MAIL" Mailing Label No. EL524961412US
Date of Deposit:_ October 25, 2000
I hereby certify that this paper or fee is being deposited with the U.S. Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Box Patent Application, Washington, D.C. 20231

Type or Print Name: CAROL MARSTALLER
Carol Marstaller
Signature

# GENE MINING SYSTEM AND METHOD

## CROSS-REFERENCES TO RELATED APPLICATIONS

This Application for Patent claims the benefit of priority from, and hereby incorporates by reference the entire disclosure of, co-pending U.S. Provisional Application for Patent Serial No. 60/161,527, filed October 26, 1999; and Serial No. 60/161,571, filed October 26, 1999.

## TECHNICAL FIELD OF THE INVENTION

This invention relates to the targeted isolation of biologically and functionally relevant gene and genomic information and bioinformatics and more particularly to a

system, method and apparatus for targeting and cloning gene sequences based on functional observations from data mined from available gene databases.

## BACKGROUND OF THE INVENTION

Without limiting the scope of the invention, its background is described in connection with uses of functional genomics and bioinformatics, as an example.

The present invention relates generally to methods and systems for searching and identifying functional nucleic acid sequences and proteins encoded by genes available from the multitude of nucleic acid and protein databases presently available. These biological databases store information that is searchable and from which biological information may be retrieved. More particularly, the present invention relates to systems and methods for identifying biologically relevant sequences of biological molecules using an integrated approach that specifically identifies sequences for cloning.

Generally, informatics may be defined as the study and application of computer and statistical techniques to the management of information. In projects related to biological information, the term "bioinformatics" has been coined to include the development of methods to, e.g., search

databases, analyze nucleic acid sequence information, predict protein sequence, protein structure, and protein function from nucleic acid sequence data.

The widespread use and availability of molecular biological techniques have allowed for the rapid development and identification of nucleic acid derived sequences. With the widespread availability of advanced computer systems and the integration of laboratory equipment with computer software, researchers are able to conduct advanced quantitative analyses, database comparisons and computational algorithms to seek and identify gene sequences with homology to known sequences.

Examples of large-scale sequencing and the availability of genetic information for a number of organisms have been cataloged in a number of public and private computer databases. Genetic databases for organisms such as *Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, and *Mycoplasma pneumoniae*, to name a few, are publicly available. At present, however, complete sequence data is available for relatively few species, and the ability to manipulate sequence data within and between species and databases is greatly limited by the ability of these public databases to be searched for functional significance.

One example of a system for comparing relational databases of sequences is disclosed in United States Patent No. 5,966,712, issued to Sabatini, et al. The system disclosed is a relational database system for storing and manipulating biomolecular sequence information and includes a database of genomic libraries for a plurality of types of organisms. These libraries are taught to have multiple genomic sequences, at least some of which represent open reading frames located along a contiguous sequence in each of the plurality of organisms' genomes. A user interface is provided and is capable of receiving a selection of two or more of the genomic libraries for comparison and displaying the results of the comparison. The system also provides a user interface capable of receiving a selection of one or more probe open reading frames for use in determining homologous matches between such probe open reading frame(s) and the open reading frames in the genomic libraries, and displaying the results of the determination.

Also needed are fully integrated systems that take advantage of functional observations and the identification of biologically relevant and functional gene sequences. This disconnect between genotype and phenotype leads to the pursuit of many genes of doubtful relevance or even mere

artifacts. Thus, researchers are presently unable to avoid using available computer resources to explore, identify and study relevant gene sequences, gene expression, and molecular structure without extensive experimentation.

Another such use of bioinformatics involves studying an organism's genome to determine the sequence and placement of its genes and their relationship to other sequences and genes within the genome or to genes in other organisms. The study of the relationship between introns and exons, for example across species, allows for a scientific understanding of many underlying substructures of the protein or proteins being expressed. It also allows for the identification of sequences that are involved in the regulation of the gene or genes that are at a particular gene locus. Such information may be of significant interest in biomedical and pharmaceutical research to assist in the evaluation of potential drug efficacy and resistance for genes that are well studied and for which significant structure-function studies have been conducted. In one such database system (Incyte Pharmaceuticals, Inc., U.S.A.), software has been developed that searched the annotated information that is part of genomic sequence data in publicly available sequence databases. Unfortunately, not all electronically recorded

sequences contain annotated information. Some contain information that is not functional, contain information that is not accurate, or contain information that has no relation to function. Examples of such databases include the widely available public databases GenBank (NCBI) and TIGR. Therefore, the accuracy and relevance of any search results from these databases often has no bearing on the cellular biological function of a particular protein of gene regulatory element.

Although genetic data processing and relational database systems such as those developed by Incyte Pharmaceuticals, Inc. provide great power and flexibility in analyzing genetic information, this area of technology is still in its infancy and further improvements in genetic data processing and relational database systems will help accelerate biological research for numerous applications.

**SUMMARY OF THE INVENTION**

While publicly available databases make manipulation of gene and genomic information easy to perform and understand, sophisticated computer database systems have not been developed that begin their searching based on functional biologically-relevant information. Furthermore, a need has

been recognized for the identification, isolation and cloning of biologically relevant genes and genomic information mined from available resources. While large amounts of sequence data are being generated as part of the Human Genome Project and other like projects, a coordinated system and method for culling functionally relevant sequences is needed. Also needed are systems and methods for mining genes based on the observation of biologic data, for which an understanding of the genetic basis for the observation is known or unknown.

The present invention provides a method for targeting gene sequences having one or more genotypic or phenotypic characteristics using a computer. One or more genotypic or phenotypic characteristics are selected. A gene sequence is then selected that is known to have the selected phenotypic characteristics. In addition one or more databases containing cataloged gene sequences are selected. The selected gene sequence is compared to the cataloged gene sequences, and any cataloged gene sequences that contain a portion of the selected gene sequence are extracted. The selected gene sequence is aligned to each portion of the extracted gene sequence and the extracted gene sequences are prioritized based on the alignment of the selected gene sequence. At least one of the prioritized gene sequences is

selected based on one or more phenotypic criteria. Finally, one or more degenerate primers are designed to target the selected-prioritized gene sequences.

The present invention also provides a computer program embodied on a computer-readable medium that performs the steps described above. In addition, the present invention provides a system having a computer, one or more databases containing the cataloged gene sequences, and a communication link connecting the computer to the one or more databases. The computer is used to select one or more phenotypic characteristics, select a gene sequence that is known to have the selected phenotypic characteristics, compare the selected gene sequence to the cataloged gene sequences, extract any cataloged gene sequences that contain a portion of the selected gene sequence, align the selected gene sequence to each portion of the extracted gene sequence, prioritize the extracted gene sequences based on the alignment of the selected gene sequence, select at least one of the prioritized gene sequences based on one or more phenotypic criteria, and design one or more degenerate primers to target the selected-prioritized gene sequences.

Thus, the present invention takes the current state of the art, which requires combing GenBank with individual

sequences to discover all of the homologous sequence, to a fully automated system that includes not only sequence parameters in the search, but includes other search parameters like species, protein characteristics and functional domains. Further, multiple homology search algorithms are seamlessly incorporated into the method. This not only allows nucleotide or amino acid searches to be performed, but allows any conceivable type of search algorithm to be employed without requiring the user to do more than select the desired parameters. In this way, multiple types of databases (e.g., nucleotide, amino acid, 3D structure, etc.) can be searched, even simultaneously if desired.

**BRIEF DESCRIPTION OF THE DRAWINGS**

For a more complete understanding of the features and advantages of the present invention, reference is now made to the detailed description of the invention along with the accompanying figures in which corresponding numerals in the different figures refer to corresponding parts and in which:

FIGURE 1 is a block diagram showing some features of the present invention;

FIGURE 2 is a basic flow chart showing a gene sequence targeting program in accordance with the present invention;

FIGURE 3 is a flow chart showing the phenotypic characteristic selection process in accordance with the present invention;

FIGURE 4 is a flow chart showing the gene sequence selection process in accordance with the present invention;

FIGURE 5 is a flow chart showing the database selection process in accordance with the present invention;

FIGURE 6 provides the system network overview in the SPADE™ system;

FIGURE 7 provides the program flow in the SPADE™ system;

FIGURE 8 provides the database management screen in the SPADE™ system;

FIGURE 9 provides the workspace management screen in the SPADE™ system;

FIGURE 10 provides the search analysis tools screen in the SPADE™ system;

FIGURE 11 provides the system architecture overview of the SPADE™ system;

FIGURE 12 provides an example of an application of the SPADE™ system;

FIGURE 13 provides an example of an application of the SPADE™ system; and

FIGURE 14 is the nucleic acid and protein sequence of an INTEGRIN protein isolated using the present invention.

**DETAILED DESCRIPTION OF THE PRESENTLY PREFERRED EXEMPLARY EMBODIMENTS**

The present invention will now be described more fully hereinafter with reference to the accompanying drawings, in which preferred embodiments of the invention are shown. This invention may, however, be embodied in many different forms and should not be construed as limited to the embodiments set forth herein; rather, these embodiments are provided so that this disclosure will be thorough and complete, and will fully convey the scope of the invention to those skilled in the art.

While the making and using of various embodiments of the present invention are discussed in detail below, it should be appreciated that the present invention provides many applicable inventive concepts that may be embodied in a wide variety of specific contexts. The specific embodiments discussed herein are merely illustrative of specific ways to

make and use the invention and do not delimit the scope of the invention.

**DEFINITIONS**

As used throughout the present specification the following abbreviations are used: TF, transcription factor; ORF, open reading frame; kb, kilobase (pairs); UTR, untranslated region; kD, kilodalton; PCR, polymerase chain reaction; RT, reverse transcriptase.

The term "x% homology" refers to the extent to which two nucleic acid or protein sequences are complementary as determined by BLAST homology alignment as described by T.A. Tatusova & T.L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS MICROBIOL LETT. **174**:247-250 and using the following parameters: Program (blastn) or (blastp) as appropriate; matrix (OBLOSUM62), reward for match (1); penalty for mismatch (-2); open gap (5) and extension gap (2) penalties; gap x- drop off (50); Expect (10); word size (11); filter (off). An example of a web based two sequence alignment program using these parameters is found at http://www.ncbi.nlm.nih.gov/gorf/bl2.html.

The invention thus includes nucleic acid or protein sequences that are highly similar to the sequences of the

present invention, and include sequences of 80, 85, 90, 95 and 98% similarity to the sequences described herein.

The invention also includes nucleic acid sequences that can be isolated from genomic or cDNA libraries or prepared synthetically, that hybridize under high stringency to the entire length of a 400 nucleotide probe derived from the nucleic acid sequences described herein under. High stringency is defined as including a final wash of 0.2X SSC at a temperature of 60°C. Under the calculation:

$$\text{Eff } T_m = 81.5 + 16.6(\log M\ [Na+]) + 0.41(\%G+C) - 0.72(\%\ \text{formamide})$$

the percentage allowable mismatch of a gene with 50% GC under these conditions is estimated to be about 12%.

The nucleic acid and protein sequences described herein are listed for convenience as follows:

| | |
|---|---|
| SEQ ID NO.: 1 | integrin beta 1 (INTB1) cDNA sequence from *M. sexta* (*see* FIGURE 14) |
| SEQ ID NO.: 2 | ITGB1 protein sequence for *M. sexta* (*see* FIGURE 14) |
| SEQ ID NO.: 3 | ITGB1 forward *primer 741-781 AAY TTG GAY WMT CYH GAR GGW GGY TTB GAT GCY MTH ATG CA |

| | |
|---|---|
| SEQ ID NO.: 4 | ITGB1 reverse primer 2358-2339 TCR AAY TTR GCA WAY TCC CT |
| SEQ ID NO.: 5 | ITGB1 forward primer 3'-RACE ATC ATT CAA ACG GAA CCA GAG |
| SEQ ID NO.: 6 | ITGB1 REV 5'-RACE GTC TCC ACC CTA TTT CTT TCT CAC |
| SEQ ID NO.: 7 | ITGB1 forward primer for sequencing TTG TGA CGG GAC ACC AAT TA |
| SEQ ID NO.: 8 | ITGB1 reverse primer for sequencing GCA TAC ACA TTC ACC GTT GC |

*Other primers used included commercially available primers from the Clontech SMART™ cDNA Library Construction Kit (SMART III Oligonucleotide; 5' PCR Primer; CDS III/3' PCR Primer; CDS III/3' TRUN).

**TOOLS**

Alignment tools for use with the present invention may include, e.g., BLAST. BLAST (Basic Local Alignment Search Tool) is a heuristic search algorithm employed by the programs blastp, blastn, blastx, tblastn, and tblastx. This combination of programs use the statistical methods of Karlin and Altschul (1990, 1993). More recent versions of the program allow for tailoring of the sequence similarity during a searching, e.g., to identify homologs in a query sequence.

The programs are not generally useful for motif-style searching.

The fundamental unit of BLAST algorithm output is the High-scoring Segment Pair (HSP). An HSP includes two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score. A set of HSPs is thus defined by two sequences, a scoring system, and a cutoff score. This HSP set may be empty if the cutoff score is sufficiently high. In the software implementation of the BLAST algorithm, each HSP has a segment from the query sequence and one from a database sequence. The sensitivity and speed of the programs may be adjusted using the standard BLAST algorithm parameters W, T, and X (Altschul et al., 1990). Furthermore, the selectivity of the programs may be adjusted via the cutoff score.

The approach to similarity searching taken by the BLAST programs is first to look for similar segments (HSPs) between the query sequence and a database sequence. Next, the statistical significance of any matches that were found is evaluated. Finally, those matches that satisfy a user-selectable threshold of significance are reported. The finding of multiple HSPs involving the query sequence and a

single database sequence are treated statistically in a variety of ways. Another problem with standard BLAST is that it uses the default programs devised for "Sum" statistics (Karlin and Altschul, 1993), as such, the statistical significance ascribed to a set of HSPs may be higher than that of any individual member of the set. Only when the ascribed significance satisfies the user-selectable threshold will the match be reported to the user.

The task of finding HSPs begins by identifying short words of length W in a query sequence that either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. The identification of the first short word as a location to initiate a search is one of the limitations of the BLAST search, as it identifies a first location to initiate an alignment and anchors its alignment at that location. By prefiltering sequences such that irrelevant sequences are removed, a priori, even the BLAST alignment tool may be used with the present invention. Furthermore, by pre-filtering the search sequences, open database BLAST searching is made more efficient by limiting search parameters to those that are functional rather than artifactual. Removal of artifactual sequences from the potential search pool further

aids in the location of relevant genes due to the limit of search results imposed by BLAST to 50 potential sequences. T is referred to as the neighborhood word score threshold (Altschul, et al., 1990). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score may be increased. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached.

A Maximal-scoring Segment Pair (MSP) is defined by two sequences and a scoring system and is the highest-scoring of all possible segment pairs that can be produced from the two sequences. The statistical methods described by Karlin and Altschul (1990, 1993) may be used to determine the significance of MSP scores in the limit of long sequences, under a random sequence model that assumes independent and identically distributed choices for the residues at each position in the sequences. These statistics may be modified by the filtering of the present invention to the task of

assessing the significance of HSP scores obtained from comparisons of pre-filtered potentially short, biological sequences.

The five BLAST programs described here perform the following tasks: blastp compares an amino acid query sequence against a protein sequence database; blastn compares a nucleotide query sequence against a nucleotide sequence database; blastx compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database; and tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames, also for both strands. More particularly, tblastx compares the six-frame translations of a nucleotide search query sequence against the six-frame translations of a nucleotide sequence database.

BLAST restricts the number of short descriptions of matching sequences reported to the number specified; default limit is 100 descriptions. During the alignment procedure, BLAST restricts database sequences to the number of specified high-scoring segment pairs (HSPS) that are requested and thereby limits its reporting function. The default HSP limit is 50. If more than 50 database sequences satisfy the

statistical significance threshold for reporting, BLAST only matches and reports those sequences given the greatest statistical significance.

The statistical significance threshold (EXCEPT value) for reporting matches against database sequences is 10, such that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Fractional values are acceptable.

The Cutoff score for reporting high-scoring segment pairs is calculated from the EXPECT value. HSPs are reported for a database sequence only if the statistical significance ascribed to them is equal to or greater that the HSP ascribed to a lone HSP having a score equal to the CUTOFF value. Higher CUTOFF values are more stringent, leading to fewer chance matches being reported. Typically, significance thresholds may be more intuitively managed using EXPECT.

Another function of BLAST is MATRIX. MATRIX is an alternative scoring matrix for BLASTP, BLASTX, TBLASTN and TBLASTX. The default matrix is BLOSUM62 (Henikoff &

Henikoff, 1992). The valid alternative choices include: PAM40, PAM120, PAM250 and IDENTITY. No alternate scoring matrices are available for BLASTN; specifying the MATRIX directive in BLASTN requests returns an error response. The STRAND function of BLAST restricts a TBLASTN search to just the top or bottom strand of the database sequences; or restrict a BLASTN, BLASTX or TBLASTX search to just reading frames on the top or bottom strand of the query sequence. The FILTER function of BLAST is limited to "mask off" segments of the query sequence that have low compositional complexity, as determined by the SEG program of Wootton & Federhen (Computers and Chemistry, 1993), or segments having short-periodicity internal repeats, as determined by the XNU program of Claverie is & States (Computers and Chemistry, 1993), or, for BLASTN, by the DUST program. Filtering may eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

Low complexity sequence found by a filter program is substituted using the letter "N" in nucleotide sequence

(e.g., "NNNNNNNNNNNNN") and the letter "X" in protein sequences (e.g., "XXXXXXXXX"). Users may turn off filtering by using the "Filter" option on the "Advanced options for the BLAST server" page.

Furthermore, filtering is only applied to the query sequence (or, its translation products), not to database sequences. Default filtering is DUST for BLASTN, SEG for other programs. It is not unusual, however, for nothing at all to be masked using the filter function of BLAST because filtering does not always yield an effect. Furthermore, in some cases, sequences are masked in their entirety, indicating that the statistical significance of any matches reported against the unfiltered query sequence should be suspect.

An alternative database searching engine for use with the present invention is another legacy system known as Clustal W. The Clustal W algorithm is basically the same as for Clustal V. Clustal W improves on the original Clustal V program, by eliminating terminal gap penalization, thereby treating them the same as all other gaps. By freeing the calculation of terminal gaps the alignment is improved by eliminating single residues jumping to the edge of the alignment.

The change in alignment scheme, however, is not without caveats, namely that a gap near the end of the alignment causes Clustal W to insert a gap thereby reducing the alignment score. By freeing terminal gaps, therefore, the overall score of an otherwise good alignment is reduced. In operation, the misalignment may be reduced by lowering the gap opening and reducing the extension penalties. It is difficult, however, to weight the balance between these two functions. The pre-filtering function of the present invention allows the user to eliminate the need to determine which of the alignment penalties to conform to by reducing the need to penalize otherwise good alignments. The present invention allows for maximum specificity and selectivity to be applied to pre-screened or filtered sequences.

One great advantage of the Clustal W program is the speed of the initial pairwise alignments. The speed of the alignment in all programs, including BLAST and others, is always commensurate with a decrease in specificity. Therefore, alignment quality is compromised for speed. Clustal W allows for a slower search speed that increases the accuracy of the alignment. By default, the initial pairwise alignments of Clustal W are carried out using a full dynamic programming algorithm. This initial pairwise alignment

is more accurate than the older hash/ k-tuple based alignments (Wilbur and Lipman) but is somewhat slower. On a fast workstation the difference in speed is often not noted. When searching larger and larger databases or clusters of databases, however, the improved filtering and searching system of the present invention greatly increases both accuracy and speed.

Another option of Clustal W is the ability to delay the alignment of distant sequences. The user may set a cut-off to delay the alignment of the most divergent sequences in a data set until all other sequences have been aligned. This delay in distant alignment is particularly useful when screening genomic sequences and is important when assessing the intron/exon junctions and intron repeats across species lines. In Clustal W the default is set to 40%, which means that if a sequence is less than 40% identical to any other sequence, its alignment will be delayed.

Clustal W also allows for the iterative realignment and for resetting gaps between alignments. By default, the alignment of a set sequences a second time (e.g., with changed gap penalties), causes the gaps from the first alignment to be discarded. Discarding the older gaps from previous alignment often provides a better alignments by

keeping the gaps (do not reset them) and doing the full multiple alignment a second time. Sometimes, the alignment will converge on a better solution, alternatively, it is possible for the new alignment will be the same as the first.

Clustal W also allows for sequence profile alignments. By profile alignment, it is meant the alignment of old alignments/sequences. In this context, a profile is just an existing alignment (or even a set of unaligned sequences). The use of a profile alignment allows the user to read in an old alignment (in any of the allowed input formats) and align one or more new sequences to that profile. The profile alignment may be a full alignment or a single sequence alignment. In the simplest mode, the user simply aligns the two profiles to each other. This cross-profile alignment is useful if to gradually build up a full multiple alignment.

A second option is to align the sequences from, e.g., a second profile, one at a time to the first profile. This is done by taking into account the underlying sequence comparison tree between the sequences. The second profile alignment is useful if the user has a set of new sequences (not aligned) and wished to add them all to an older alignment.

Examples of databases that may be used to prescreen for sequences include both public and private databases of either nucleic acid or protein sequences. As will be understood by those of skill in the art, nucleic acids generally may be either ribonucleic acids or deoxyribonucleic acids, or derivatives or variants thereof.

One such database is ACEDB. Acedb is a genome database system developed over the last 7 years primarily by Jean Thierry-Mieg (CNRS, Montpellier) and Richard Durbin (Sanger Centre). It provides a custom database kernel, with a non-standard data model designed specifically for handling scientific data flexibly and a graphical user interface with many specific displays and tools for genomic data.

Acedb may be used for both managing data within genome projects, and for making genomic data available to other scientists. Acedb was originally developed for the C.elegans genome project, from which its name was derived (A C.elegans DataBase). The tools in it have been generalized to allow for greater flexibility to the point that the same software is now used for many different genomic databases from, e.g., bacteria, fungi, plants to man. It is also increasingly used for databases with non-biological content, e.g., vectors and viruses.

The acedb software is primarily developed to run under the Unix operating system, using X-Windows for graphics. Copies of the software are accessible via FTP sites, or may be interfaced with through a Web interface, which serves a number of human databases as well as the AceBrowser system, which serves a local installation of the C.elegans Genome Database.

Referring to FIGURE 1, a block diagram shows some features of the present invention. The gene sequence targeting program 100 of the present invention comprises a variety of tool types, such as interface tools 110, targeting tools 120, analysis tools 130, design tools 140, and cloning tools 150. These tools 110, 120, 130, 140 and 150 are preferably integrated together using an objected-oriented programming language.

The interface tools 110 may include a graphical user interface (GUI) 112, one or more interfaces with public and private databases 114, and data storage and output tools 116. The GUI 112 is preferably a menu driven interface that allows a user to jump between applications, point and click on selections, and view information in graphical form. The one or more interfaces with public and private databases 114 allow the program and the user to access, search and retrieve

data from local and remote databases, which may be public or private. These interfaces 114 can be conFIGUREd to allow seamless access to a variety of disparate databases, such as publication databases and gene sequence databases. The data storage and output tools 116 may provide access to program help information, experimental documentation features, reports, project data storage, and data backup, import and export features.

The following sequence comparison software is available from the Genetics Computer Group (GCG) software and may be accessed by the system of the present invention.

**TABLE I SEQUENCE RETRIEVAL-INTERFACE TOOLS**

**Fetch**

Copies GCG sequences or data files from the GCG database into your directory or displays them on your terminal screen.

**NetFetch**

Retrieves entries from NCBI listed in a NetBLAST output file. It can also be used to retrieve entries individually by entry name or accession number. The output of NetFetch is an RSF file.

The targeting tools 120 allow the user to set the parameters that will be used to target the gene sequence.

These targeting tools 120 may include a phenotypic characteristics selection process 122, a gene process 124 and a database selection process 126. The phenotypic characteristics selection process 122, gene selection process 124 and database selection process 126 will be described below in more detail in reference to FIGURES 3, 4 and 5 respectively.

The following database searching software is available from the Genetics Computer Group (GCG) software and may be accessed by the system of the present invention.

**TABLE II DATABASE SEARCHING-TARGETING TOOLS**

**Reference Searching**

**LookUp**

Identifies sequence database entries by name, accession number, author, organism, keyword, title, reference, feature, definition, length, or date. The output is a list of sequences.

**StringSearch**

Identifies sequences by searching for character patterns such as "globin" or "human" in the sequence documentation.

**Names**

Identifies GCG® data files and sequence entries by name. It may show what set of sequences is implied by any sequence specification.

The analysis tools 130 generate results based on the information and preferences selected by user with the targeting tools 120 and then allow the user to analyze those results. The analysis tools 130 may include a comparison and extraction process 132, an alignment process 134 and a prioritizing and filtering process 136. These analysis tools 130 can be legacy systems.

The following analysis tools software is available from the Genetics Computer Group (GCG) software and may be accessed by the system of the present invention.

**TABLE III MULTIPLE SEQUENCE COMPARISON-ANALYSIS TOOLS**

**Gap**

Uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps.

**BestFit**

Makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the local homology algorithm of Smith and Waterman.

**FrameAlign**

Creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in all possible reading frames on a single strand of a nucleotide sequence. Optimal alignments may include reading frame shifts.

**Compare**

Compares two protein or nucleic acid sequences and creates a file of the points of similarity between them for plotting with DotPlot. Compare finds the points using either a window/stringency or a word match criterion. The word comparison is 1,000 times faster than the window/stringency comparison, but somewhat less sensitive.

**DotPlot**

Makes a dot-plot with the output file from Compare or StemLoop.

**GapShow**

Displays an alignment by making a graph that shows the distribution of similarities and gaps. The two input sequences should be aligned with either Gap or BestFit before they are given to GapShow for display.

**ProfileGap**

Makes an optimal alignment between a profile and one or more sequences.

**Pileup**

Creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It may also plot a tree showing the clustering relationships used to create the alignment.

**PlotSimilarity**

Plots the running average of the similarity among the sequences in a multiple sequence alignment.

**MEME**

(Multiple EM for Motif Elicitation) Finds motifs in a group of unaligned sequences. MEME saves these motifs as a set of profiles. A database search of sequences with these profiles is then conducted using, e.g., the MotifSearch program.

**ProfileMake**

Creates a position-specific scoring table, called a profile, that quantitatively represents the information from a group of aligned sequences. The profile may then be used for database searching (ProfileSearch) or sequence alignment (ProfileGap).

**ProfileGap**

Makes an optimal alignment between a profile and one or more sequences.

**Overlap**

Compares two sets of DNA sequences to each other in both orientations using a WordSearch style comparison.

**NoOverlap**

Identifies the places where a group of nucleotide sequences do not share any common subsequences.

**OldDistances**

Makes a table of the pairwise similarities within a group of aligned sequences.

TABLE IV DATABASE SEARCHING-ANALYSIS TOOLS

**Sequence Searching**

**BLAST**

Searches for sequences similar to a query sequence. The query and the database searched may be either peptide or nucleic acid in any combination. BLAST can search databases on a local computer or databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA.

**NetBLAST**

Searches for sequences similar to a query sequence. The query and the database searched may be either peptide or nucleic acid in any combination. NetBLAST can search only databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA.

**FastA**

Does a Pearson and Lipman search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein). For nucleotide searches, FastA may be more sensitive than BLAST.

**SSearch**

Does a rigorous Smith-Waterman search for similarity between a query sequence and a group of sequences of the same

type (nucleic acid or protein). This may be the most sensitive method available for similarity searches. Compared to BLAST and FastA, it is very slow.

**TFastA**

Does a Pearson and Lipman search for similarity between a protein query sequence and any group of nucleotide sequences. TfastA translates the nucleotide sequences in all six reading frames before performing the comparison. It is designed to answer the question, "What implied protein sequences in a nucleotide sequence database are similar to my protein sequence?"

**TFastX**

Does a Pearson and Lipman search for similarity between a protein query sequence and any group of nucleotide sequences, taking frameshifts into account. It is designed to be a replacement for TfastA, and like TfastA, it is designed to answer the question, "What implied protein sequences in a nucleotide sequence database are similar to my protein sequence?"

**FastX**

Does a Pearson and Lipman search for similarity between a protein query sequence and any group of nucleotide sequences. TfastA translates the nucleotide sequences in all

six reading frames before performing the comparison. It is designed to answer the question, "What implied protein sequences in a nucleotide sequence database are similar to my protein sequence?"

**FrameSearch**

Searches a group of protein sequences for similarity to one or more nucleotide query sequences, or searches a group of nucleotide sequences for similarity to one or more protein query sequences. For each sequence comparison, the program finds an optimal alignment between the protein sequence and all possible codons on each strand of the nucleotide sequence. Optimal alignments may include reading frame shifts.

**MotifSearch**

Uses a set of profiles (representing similarities within a family of sequences) as a query to either a) search a database for new sequences similar to the original family, or b) annotate the members of the original family with details of the matches between the profiles and each of the members. Normally, the profiles are created with the program MEME.

**ProfileSearch**

Uses a profile (representing a group of aligned sequences) as a query to search the database for new sequences with similarity to the group. The profile is created with the program ProfileMake.

**ProfileSegments**

Makes optimal alignments showing the segments of similarity found by ProfileSearch.

**FindPatterns**

Identifies sequences that contain short patterns like GAATTC or YRYRYRYR. Patterns may be define ambiguously, thereby allowing for a greater number of mismatches. Patterns may be provided in a file or simply typed into a terminal.

**Motifs**

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE® Dictionary of Protein Sites and Patterns. Motifs can display an abstract of the current literature on each of the motifs it finds.

**WordSearch**

Identifies sequences in the database that share large numbers of common words in the same register of comparison

with your query sequence. The output of WordSearch can be displayed with Segments.

**Segments**

Aligns and displays the segments of similarity found by WordSearch.

**LineUp**

Is a screen editor for editing multiple sequence alignments. Up to 30 sequences may be edited simultaneously. New sequences may also be typed in by hand or added from existing sequence files. A consensus sequence identifies places where the sequences are in conflict.

**TABLE V FRAGMENT ASSEMBLY-ANALYSIS TOOLS**

**GelStart**

Begins a fragment assembly session by creating a new fragment assembly project or by identifying an existing project.

**GelEnter**

Adds fragment sequences to a fragment assembly project. It accepts sequence data from your terminal keyboard, a digitizer, or existing sequence files.

**GelMerge**

Aligns the sequences in a fragment assembly project into assemblies called contigs. The assembled contigs may be viewed and/or edited from the assemblies generated in GelAssemble.

**GelAssemble**

Is a multiple sequence editor for viewing and editing contigs assembled by GelMerge.

**GelView**

Displays the structure of the contilas in a fragment assembly project.

**GelDisassemble**

Breaks up the contigs in a fragment assembly project into single fragments.

**TABLE VI GENE FINDING AND PATTERN RECOGNITION-ANALYSIS TOOLS**

**TestCode**

Helps you identify protein coding sequences by plotting a measure of the non-randomness of the composition at every third base. The statistic does not require a codon frequency table.

**CodonPreference**

Is a frame-specific gene finder that tries to recognize protein coding sequences by virtue of the similarity of their codon usage to a codon frequency table or by the bias of their composition (usually GC) in the third position of each codon.

**Frames**

Shows open reading frames for the six translation frames of a DNA sequence. Frames may superimpose the pattern of rare codon choices if you provide it with a codon frequency table.

**Terminator**

Searches for prokaryotic factor-independent RNA polymerase terminators according to the method of Brendel and Trifonov.

**Motifs**

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE© Dictionary of Protein Sites and Patterns. Motifs can display an abstract of the current literature on each of the motifs it finds.

**MEME**

(Multiple EM for Motif Elicitation) Finds conserved motifs in a group unaligned sequences. MEME saves these

motifs as a set of profiles. A database search for sequences with similar profiles may be conducted using the MotifSearch program.

**Repeat**

Finds direct repeats in sequences. You must set the size, stringency, and range within which the repeat must occur; all the repeats of that size or greater are displayed as short alignments.

**FindPatterns**

Identifies sequences that contain short patterns like GAATTC or YRYRYRYR. The user may define the patterns ambiguously and allow mismatches or provide the patterns in a file or simply type them in from the terminal.

**Composition**

Determines the composition of sequence(s). For nucleotide sequence(s), Composition also determines dinucleotide and trinucleotide content.

**CodonFrequency**

Tabulates codon usage from sequences and/or existing codon usage tables. The output file is correctly formatted for input to the CodonPreference, Correspond, and Frames programs.

**Correspond**

Looks for similar patterns of codon usage by comparing codon frequency tables.

**Window**

Makes a table of the frequencies of different sequence patterns within a window as it is moved along a sequence. A pattern is any short sequence like GC or R or ATG. The sata output may be ploted with the program StatPlot.

**StatPlot**

Plots a set of parallel curves from a table of numbers like the table written by the Window program. The statistics in each column of the table are associated with a position in the analyzed sequence.

**FitConsensus**

Uses a consensus table written by Consensus as a probe to find the best examples of the consensus in a DNA sequence. The number of fits may be specified by the user and FitConsensus tabulates them with their position, frame, and a statistical measure of their quality.

**Consensus**

Calculates a consensus sequence for a set of pre-aligned short nucleic acid sequences by tabulating the percent of G, A, T, and C for each position in the set. FitConsensus uses

the Consensus output table as a probe to search for the best examples of the derived consensus in other nucleotide sequences.

**Xnu**

Replaces statistically significant tandem repeats in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search, the regions with X characters are ignored.

**Seg**

Replaces low complexity regions in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search, the regions with X characters are ignored.

**TABLE VII PROTEIN ANALYSIS-ANALYSIS TOOLS**

**Motifs**

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE© Dictionary of Protein Sites and Patterns. Motifs can display an abstract of the current literature on each of the motifs it finds.

**ProfileScan**

Uses a database of profiles to find structural and sequence motifs in protein sequences.

**CoilScan**

Locates coiled-coil segments in protein sequences.

**HTHScan**

Scans protein sequences for the presence of helix-turn-helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.

**SPScan**

Scans protein sequences for the presence of secretary signal peptides (SPs).

**PeptideSort**

Shows the peptide fragments from a digest of an amino acid sequence. It sorts the peptides by weight, position, and HPLC retention at pH 2.1. and shows the composition of each peptide. It also prints a summary of the composition of the whole protein.

**Isoelectric**

Plots the charge as a function of pH for any peptide sequence.

**PeptideMap**

Creates a peptide map of an amino acid sequence.

**PepPlot**

Plots measures of protein secondary structure and hydrophobicity in parallel panels of the same plot.

**PeptideStructure**

Makes secondary structure predictions for a peptide sequence. The predictions include (in addition to alpha, beta, coil, and turn) measures for antigenicity, flexibility, hydrophobicity, and surface probability. PlotStructure displays the predictions graphically.

**Plotstructure**

Plots the measures of protein secondary structure in the output file from PeptideStructure. The measures may be shown on parallel panels of a graph or with a two-dimensional "squiggly" representation.

**Moment**

Makes a contour plot of the helical hydrophobic moment of a peptide sequence.

**HelicalWheel**

Plots a peptide sequence as a helical wheel to help you recognize amphiphilic regions.

**Xnu**

Replaces statistically significant tandem repeats in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search, the regions with X characters are ignored.

**Seg**

Replaces low complexity regions in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search, the regions with X characters are ignored.

The design tools 140 allow the user to select a gene sequence and design degenerate primers.

The design tools 140 may include a gene sequence selection process 142 and a degenerate primer design process 144. The following analysis tools software is available from the Genetics Computer Group (GCG) software and may be accessed by the system of the present invention.

**TABLE VIII PRIMER SELECTION-DESIGN TOOLS**

**Prime**

Selects oligonucleotide primers for a template DNA sequence. The primers may be useful for the polymerase chain reaction (PCR) or for DNA sequencing. Prime allows the user to choose primers from the whole template or limit the choices to a particular set of primers listed in a file.

**TABLE IX EVOLUTION-DESIGN TOOLS**

**PAUPSearch**

Provides a GCG interface to the tree-searching options in PAUP (Phylogenetic Analysis Using Parsimony). Starting with a set of aligned sequences, a search may be conducted for phylogenetic trees that are optimal ccording to parsimony, distance, or maximum likelihood criteria; reconstruct a neighbor-joining tree; or perform a bootstrap analysis.

**Distances**

Creates a table of the pairwise distances within a group of aligned sequences.

**GrowTree**

Creates a phylogenetic tree from a distance matrix created by Distances using either the UPGMA or neighbor-joining method. A text or graphics output file may be conducted.

**Diverge**

Estimates the pairwise number of synonymous and nonsynonymous substitutions per site between two or more aligned nucleic acid sequences that code for proteins.

The cloning tools 150 allow the user to clone genetic material from the degenerate primers via cloning process 152 as described hereinbelow in the examples.

Now referring to FIGURE 2, a basic flow chart shows a gene sequence targeting program 200 in accordance with the present invention. The gene sequence targeting program 200 begins in block 202. One or more phenotypic characteristics are selected using the phenotypic characteristic selection process (see FIGURE 3) in block 204. A gene sequence that is known to have the selected phenotypic characteristics is selected using the gene sequence selection process (see FIGURE 4) in block 206. One or more databases containing cataloged gene sequences are selected using the database selection process (see FIGURE 5) in block 208.

The selected gene sequence is compared to the cataloged gene sequences in block 210, and any cataloged gene sequences that contain a portion of the selected gene sequence are extracted in block 212. The selected gene sequence is aligned to each portion of the extracted gene sequence in block 214 and the extracted gene sequences are prioritized and filtered based on the alignment of the selected gene sequence in block 216. At least one of the prioritized gene sequences is selected based on one or more phenotypic criteria in block

218. One or more degenerate primers are designed to target the selected-prioritized gene sequences in block 220, and genetic material is cloned using the one or more degenerate primers in block 222. The program is complete in block 224.

Referring now to FIGURE 3, a flow chart shows the phenotypic characteristic selection process 204 in accordance with the present invention. The phenotypic characteristic selection process 204 begins in block 302 and a list of available phenotypic characteristics is displayed to the user via the GUI 112 (FIGURE 1) in block 304. The user can select one of the displayed phenotypic characteristics, read one or more phenotypic characteristics from storage, such as a data file, or create a new phenotypic characteristic selection option. If the user selects the option of picking one of the displayed phenotypic characteristics, as determined in decision block 306, the selected phenotypic characteristic is read in block 308. The user is then prompted to select additional phenotypic characteristics in block 310.

If the user selects the option of reading one or more phenotypic characteristics from storage, as determined in decision block 306, the user identifies the location of the stored data in block 314. The location of the stored data may be accessed locally via a disk drive or remotely via a

network. The phenotypic characteristics are then read from storage in block 316. Standard error handling routines can be used to report status of the read operation, test the data, prompt the user for additional information, or indicate that the read was not successfully completed. The user is then prompted to select additional phenotypic characteristics in block 310.

If the user selects the option of creating a new phenotypic characteristic selection option, as determined in decision block 306, the new phenotypic characteristic data is read in block 318. This new data can be entered directly by the user or read from a file. The new phenotypic characteristic data is stored in block 320 and can be included in the list of available phenotypic characteristics displayed in block 304. If the new phenotypic characteristic data has errors or was not properly read and stored, as determined in decision block 322, the error is reported in block 324. If a maximum number of retry attempts has not occurred, as determined in decision block 326, the new characteristic process repeats by again reading the new phenotypic characteristic data in block 318. If, however, there are no errors, as determined in decision block 322, or the maximum number of retry attempts has occurred, as

determined in decision block 326, the user is prompted to select additional phenotypic characteristics in block 310.

After the selected method is complete (see blocks 308, 316, 322 and 326), the user may then elect to select additional phenotypic characteristics. If the user elects to select additional phenotypic characteristics, as determined in is decision block 310, the list of available phenotypic characteristics is displayed again in block 304 and the process repeats as previously described. If, however, the user elects to not select additional phenotypic characteristics, as determined in decision block 310, processing returns to the main program in block 312.

Now referring to FIGURE 4, a flow chart shows the gene sequence selection process 206 in accordance with the present invention. The gene selection process 206 begins in block 402. The user can enter a gene sequence using the GUI, read a gene sequence from storage, such as a data file, or search for all or part of a gene sequence. If the user selects the option of entering a gene sequence using the GUI, as determined in decision block 404, the gene sequence is read in block 406 and processing returns to the main program in block 408.

If the user selects the option of reading a gene sequence from storage, as determined in decision block 404, the user identifies the location of the stored data in block 410. The location of the stored data may be accessed locally via a disk drive or remotely via a network. The gene sequence is then read from storage in block 412 and processing returns to the main program in block 408. Standard error handling routines can be used to report status of the read operation, test the data, prompt the user for additional information, or indicate that the read was not successfully completed.

If the user selects the option of searching for all or part of a gene sequence, as determined in decision block 404, the search parameters, such as the database to be searched, are defined in block 414. The search is performed in block 416. If a gene sequence was not found, as determined in decision block 418, the user is again prompted to select a gene sequence selection method in block 404. If, however, a gene sequence was found, as determined in decision block 418, the search results are displayed in block 420. The user can then run a new search, save the search results, select a gene sequence from the search results or exit the selection process. If the user elects to run a new search, as

determined in decision block 422, processing returns to block 414 where the search parameters are again defined. If the user elects to save the search results, as determined in decision block 422, the search results are then save to storage in block 424 and the user can then run a new search, save the search results, select a gene sequence from the search results or exit the selection process. If the user elects to select a gene sequence from the search results, as determined in decision block 422, the gene sequence is selected in block 426 and the user can then run a new search, save the search results, select a gene sequence from the search results or exit the selection process. If the user elects to exit the process, as determined in decision block 422, processing returns to the main program in block 408.

Referring now to FIGURE 5, a flow chart shows the database selection process 208 in accordance with the present invention. The database selection process 208 begins in block 502 and a list of available databases is displayed to the user via the GUI 112 (FIGURE 1) in block 504. The user can select one of the displayed databases, or provide the necessary information to search a new database. If the user selects the option of picking one of the displayed databases, as determined in decision block 305, the database selection

is read in block 508. A list of available superfamilies, families and subfamilies for the selected database is displayed in block 510 and the family selection is read in block 512. The user is then prompted to select additional databases in block 514.

If the user selects the option of providing the necessary information to search a new database, as determined in decision block 506, the data necessary to read the new database is read in block 518. This new data can be entered directly by the user or read from a file. The new database information is stored in block 520 and can be included in the is list of available databases displayed in block 504. If the new database information has errors or was not properly read and stored, as determined in decision block 522, the error is reported in block 524. If a maximum number of retry attempts has not occurred, as determined in decision block 526, the new database process repeats by again reading the information necessary to search the new database in block 518, if, however, there are no errors, as determined in decision block 522, or the maximum number of retry attempts has occurred, as determined in decision block 526, the user is prompted to select additional databases in block 514.

Afer the selected method is complete (see blocks 512, 522 and 526), the user may then elect to select additional databases. If the user elects to select additional databases, as determined in decision block 514, the list of available databases is displayed again in block 504 and the process repeats as previously described. If, however, the user elects to not select additional databases, as determined in decision block 514, processing returns to the main program in block 516.

It should be understood that all of the above processes are capable of being executed either on a single computer, or via a coordinating network of computers, each of which is capable of executing any of the described processes. It should further be understood that the invention set forth herein may be stored within computer memory, or on a hard drive or multiple hard drives of one or more computers, server or other media, e.g., CD-ROM or diskette.

A system of data mining tools has been developed to help identify, isolate and clone biologically and functionally important genes from public genomic libraries. The software suite called SPADE™, is designed to seamlessly integrate available search and analysis tools so that computer experiments for sequence analysis can be quickly designed and

executed and that rational primer design, cloning and protein characterization can be accomplished.

SPADE™ is a client/server application. The clients interact with the server, which can be a dedicated LINUX server, via a local area network or a web interface. Therefore, the interaction is platform-free. An example of the system network overview is illustrated in FIGURE 6.

An illustration of the main program flow is exemplified in FIGURE 7. A user first logs in and is the presented with a main menu. The main menu presents four choices: Database Management (FIGURE 8), Workspace Management (FIGURE 9), Search Tools and Analysis Tools (FIGURE 10). The Database Management screen allows the administrator of the system to conFIGURE the local genomic databases associated with SPADE™. In this screen, there is a list of current databases online, a button to edit the configuration for each individual database, and options to add new databases or delete existing existing databases. The Workspace Management screen allows the user to access his or her data, files and documentation on the server. It is similar to a file management program. There is a list of projects, and the files in the current project. The user can open a project, create new projects or delete existing projects. Within each

project, the user can open individual data files, rename, delete, upload or download files. The search tool screen allows the user to search databases with the algorithms associated with SPADE™. The user first selects the database via a database selection window, and then selects the sequence to search from the project files or enters the sequence directly into the text box. The user then selects the algorithm to search, and accepts the default parameters or modifies the appropriate parameters. Users can access the advance parameters via the advance parameters screen. Finally, the server executes the search and returns the result to the user. The search tool screen also allows the user to analyze the results of the previous search or analysis with the algorithms associated with SPADE™. The user first selects the sequence to analyze from the project files or enters the sequence directly into the text box. The user then selects the algorithm to execute, and accepts the default parameters or modifies the appropriate parameters. Users can access the advance parameters via the advance parameters screen. Finally, the server executes the algorithm and returns the result to the user.

An example of the system architecture overview is illustrated in FIGURE 11, showing the interaction of the

platform-free users with the four screens discussed above. FIGURE 12 describes a use of the system described in FIGURE 11. A more specific example of the application is outlined in FIGURE 13, which shows one possible use of the SPADE™ system.

The seamless integration of the various components described in the process flow discussed above, allows for the modification of existing components and the introduction of additional components which facilitate the characterization, targeting, cloning, validation, search and analysis, sorting, indexing, cataloging and conversion of various forms and formats of data and databases including, but not limited to, DNA sequences, amino acid sequences, DNA and protein motifs, images, patterns, and tertiary and quarternary structure including, atomic and molecular-level interactions. Therefore, the system described above may be used to perform high throughput database conversion, high specificity and high throughput selection of primers, as well as high specificity and high throughput positioning of protein and DNA structure and motifs. In addition, each of the various components described in the process flow discussed above may be used individually or in combination with the remaining components, thereby allowing for the delivery of results

from an individual component or a combination of components, as desired.

**EXAMPLE 1 ISOLATION OF NUCLEIC ACID MOLECULES RELATED TO INTEGRIN**

The integrin family of cell adhesion receptors plays a fundamental role in the processes involved in cell division, differentiation and movement. The extracellular domains of integrin alpha/beta heterodimers mediate cell-matrix and cell-cell contacts while their cytoplasmic tails associate with the cytoskeleton and integrins can transduce information bidirectionally. Studies have led to the identification of the ligand-binding region on the beta subunit and sequences in the cytoplasmic tails of the beta subunits that interact with cytoskeletal and signalling components. Green L.J. et al., *The integrin beta subunit*. Int J Biochem Cell Biol (1998) **30(2)**:179-84. Integrin beta 1 (ITGB1) is a subunit of type I membrane proteins and has cysteine rich domains that are involved in intrachain disulfide bonds. It associates with the alpha-1 or alpha-6 subunits to form a laminin receptor, with alpha-2 to form a collagen receptor, with alpha-4 to interact with vcam-1, with alpha-5 to form a fibronectin receptor and with alpha-8.

In order to demonstrate the system and method for identifying functional proteins in other target organisms, an integrin-like molecule most closely related to integrin beta 1 was identified and cloned from *Manduca sexta* (*M. sexta*). In this example, the original phenotypic characteristics selected were that the target molecule include a specific function and tissue localization. The specific function identified was that the target be an integral membrane protein involved in cytoskeletal formation. The localization selected was that the protein be expressed in the midgut of an organism.

These structural-functional parameters were then used to target potential genes based on the function identified from the PubMed database on all organisms (see FIGURE 2). That is, the original search for a protein was not restricted by filtering.

Following the initial identification of a target and the filtering of sequences, an alignment of the beta integrin proteins that were identified from all organisms was conducted and primer selection was made based on the identified matching sequences between the different organisms. The primer design software was the MacVector software, and following an initial round of sequence

determination, the primer design was improved. The exact primers used are provided in the SEQ ID Listing.

RT-PCR was conducted from *M. sexta* mRNA and following the PCR reaction a band of the expected size was cut out of a low-melt agarose gel. The PCR products were then cloned into the pAT vector and inserts sequenced. A BLAST alignment of the sequences identified a clone with similarity to *Pacifastacus leniusculus* (signal crayfish), *Drosophila* (fruit fly), *Anopheles gambiae* (African malaria mosquito) integrin beta 1 sequences.

The insert from these clones was then used to clone the full-length cDNA from a *M. sexta* library. The sequence of integrin beta 1 (ITGB1) gene is depicted in FIGURE 14 as SEQ ID NO.:1 and the corresponding amino acid sequence is at SEQ ID NO.: 2. These sequences represent preliminary sequence data, and the sequences will be completed and confirmed by methods known in the art.

The closest homology of this partial protein sequence is to the beta integrin of the fruit fly (Acc. No. A30889) at 146/379 (38%) identities and 216/379 (56%) similarities. The divergence at the carboxy end (beginning at aa 355) of the fragment may indicate that the sequence has an error,

resulting in a frame shift. Work is in progress to finalize and confirm the entire sequence of the novel gene.

**EXAMPLE 2 ISOLATION OF A KNOWN GENE TO VALIDATE SYSTEM**

In order to validate the system, it was used to isolate a known gene; in this case the *M. sexta* aminopeptidase gene. Aminopeptidase is involved in the modulation of various cellular responses, especially in cell-cell adhesion and signal transduction. We are particularly interested in aminopeptidase because we have shown that it is directly involved in resistance by insects to insecticidal toxins of Bacillus thuringiensis. We believe that it is a major factor involved in innate immunity of invertebrate and vertebrate epithelial cells. The *M. sexta* aminopeptidase gene was mined based on nucleotide and amino acid sequence alignment with the existing aminopeptidase related sequences, excluding the tobacco hornworm sequences. The primers used for PCR were based on such alignment.

Using this method, the tobacco hornworm aminopeptidase gene has been partially cloned and sequenced (not shown). The amino acid sequence fragments showed high homology (99-100%) to GenBank Acc. No. P91885 (Denolf, P. *et al.*, *Cloning and characterization of Manduca sexta and Plutella*

*xylostella midgut aminopeptidase N enzymes related to Bacillus thuringiensis toxin-binding proteins* Eur. J. Biochem. **248(3)**, 748-761 (1997)). Thus, the gene mining technique has been proven to isolate a known gene.

**EXAMPLE 3 FUTURE EXPERIMENTS**

The above insect genes will be further characterized according to well established methods. Protein and peptide antibodies are made according to established protocols. The antibodies are used to confirm tissue and cellular localization of the expressed protein. The extent of homology of the identified genes with other insect species and other genera is checked by zooblot at varying hybridization stringencies. The recombinant proteins are expressed, in for example, insect SF9 cells, and purified using the above antibodies, by GST or HIS tag immunoaffinity or by other means known in the art. The genes are mutated to prepare truncation mutants in order to delineate the boundaries of the functional proteins.

While this invention has been described in reference to illustrative embodiments, this description is not intended to be construed in a limiting sense. Various modifications and combinations of the illustrative embodiments, as well as

other embodiments of the invention, will be apparent to persons skilled in the art upon reference to the description. It is therefore intended that the appended claims encompass any such modifications or embodiments.

**WHAT IS CLAIMED IS:**

1. A purified nucleic acid molecule, comprising a nucleic acid sequence encoding SEQ ID NO.: 2.

2. The purified nucleic acid molecule of claim 1, which is a cDNA molecule.

3. The purified nucleic acid molecule of claim 2, which comprises the sequence of SEQ ID NO.: 1.

4. A purified nucleic acid, wherein said nucleic acid is capable of hybridizing at high stringency to a probe of 400 contiguous nucleotides from SEQ ID NO.: 1 over the entire length of said probe.

5. A purified nucleic acid, comprising a sequence that encodes a protein that is at least 90% homologous to the entire length of amino acid sequence of SEQ ID NO.: 2.

6. The purified nucleic acid of claim 5, wherein the protein is at least 95% homologous to SEQ ID NO.: 2.

7. The purified nucleic acid of claim 5, wherein the protein is at least 98% homologous to SEQ ID NO.: 2.

8. A purified protein, comprising a sequence that is at least 80% homologous to the entire length of SEQ ID NO.: 2.

9. The purified protein of claim 8, wherein the sequence is at least 90% homologous to SEQ ID NO.: 2.

10. The purified protein of claim 9, wherein the sequence is at least 95% homologous to SEQ ID NO.: 2.

11. The purified protein of claim 9, wherein the sequence is at least 98% homologous to SEQ ID NO.: 2.

12. The purified protein of claim 9, wherein the sequence is SEQ ID NO.: 2.

13. A method for targeting gene sequences having one or more phenotypic characteristics using a computer, the method comprising the steps of:

selecting one or more phenotypic characteristics;

selecting a gene sequence that is known to have the selected phenotypic characteristics;

selecting one or more databases containing cataloged gene sequences;

comparing the selected gene sequence to the cataloged gene sequences;

extracting any cataloged gene sequences that contain a portion of the selected gene sequence;

aligning the selected gene sequence to each portion of the extracted gene sequence;

prioritizing the extracted gene sequences based on the alignment of the selected gene sequence;

selecting at least one of the prioritized gene sequences based on one or more phenotypic criteria; and

designing one or more degenerate primers to target the selected-prioritized gene sequences.

14. The method as recited in claim 13, further comprising the step of filtering the prioritized gene sequences.

15. The method as recited in claim 14, wherein the step of filtering the prioritized gene sequences removes vertebrate sequences but not invertebrate derived sequences.

16. The method as recited in claim 13, further comprising the step of cloning genetic material using the one or more degenerate primers.

17. The method as recited in claim 13, wherein the one or more databases are selected from cataloged gene sequences for humans, rats, mice, zebra fish, frogs, Drosophila, nematode, C. elegans, mosquito and bacteria.

18. The method as recited in claim 13, wherein the phenotypic characteristics include insect mid-gut epithelial cell encoded proteins.

19. The method as recited in claim 13, wherein the one or more degenerate primers are nested.

20. The method as recited in claim 13, wherein the one or more degenerate primers is used to clone target molecules.

21. The method as recited in claim 13, wherein the one or more degenerate primers is used to clone biopesticide encoding genes.

22. The method as recited in claim 13, wherein the one or more degenerate primers is used to clone therapeutic encoding genes.

23. The method as recited in claim 13, wherein the step of prioritizing the extracted gene sequences based on the alignment of the selected gene sequence is accomplished by using a statistical analysis of the alignment.

24. The method as recited in claim 13, wherein the step of aligning the selected gene sequences to each extracted

gene sequence is accomplished using a local alignment search tool.

25. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by amino acid sequences.

26. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by nucleic acid sequences.

27. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by genomic DNA.

28. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by open reading frames.

29. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by introns.

30. The method as recited in claim 13, wherein the selected gene sequence is aligned to each extracted gene sequence by introns and exons.

31. The method as recited in claim 13, wherein the one or more phenotypic criteria excludes genes encoded by mammals.

32. The method as recited in claim 13, wherein the one or more phenotypic criteria excludes genes encoded by zebra fish or frogs.

33. The method as recited in claim 13, wherein the one or more phenotypic criteria excludes genes encoded by invertebrates.

34. A system for targeting gene sequences having one or more characteristics comprising:

a computer having program means thereon for selecting one or more phenotypic characteristics, selecting a gene sequence that is known to have the selected phenotypic characteristics, comparing the selected gene sequence to the cataloged gene sequences, extracting any cataloged gene

sequences that contain a portion of the selected gene sequence, aligning the selected gene sequence to each portion of the extracted gene sequence, prioritizing the extracted gene sequences based on the alignment of the selected gene sequence, selecting at least one of the prioritized gene sequences based on one or more phenotypic criteria, and designing one or more degenerate primers to target the selected-prioritized gene sequences;

one or more databases containing the cataloged gene sequences; and

a communication link connecting the computer to said one or more databases.

35. The system as recited in claim 34, further comprising:

at least one other computer, connected to said computer, said at least one other computer having said program means thereon for selecting one or more phenotypic characteristics, selecting a gene sequence that is known to have the selected phenotypic characteristics, comparing the selected gene sequence to the cataloged gene sequences, extracting any cataloged gene sequences that contain a portion of the selected gene sequence, aligning the selected gene sequence

to each portion of the extracted gene sequence, prioritizing the extracted gene sequences based on the alignment of the selected gene sequence, selecting at least one of the prioritized gene sequences based on one or more phenotypic criteria, and designing one or more degenerate primers to target the selected-prioritized gene sequences.

36. The system as recited in claim 34 or 35, wherein the program means on said computer filters the prioritized gene sequences.

37. The system as recited in claim 36, wherein the program means on said computer removes vertebrate sequences but not invertebrate derived sequences when the prioritized sequences are filtered.

38. The system as recited in claim 36, further comprising an apparatus that clones genetic material using one or more degenerate primers.

39. The system as recited in claim 36, wherein the one or more databases are selected from cataloged gene sequences

for humans, rats, mice, zebra fish, frogs, Drosophila, nematode, C. elegans, mosquito and bacteria.

40. The system as recited in claim 36, wherein the phenotypic characteristics include insect mid-gut epithelial cell encoded proteins.

41. The system as recited in claim 36, wherein the one or more degenerate primers are nested.

42. The system as recited in claim 36, wherein the one or more degenerate primers is used to clone target molecules.

43. The system as recited in claim 36, wherein the one or more degenerate primers is used to clone biopesticide encoding genes.

44. The system as recited in claim 36, wherein the one or more degenerate primers is used to clone therapeutic encoding genes.

45. The system as recited in claim 36, wherein the program means on said computer uses a statistical analysis

of the alignment of the selected gene sequence to prioritize the extracted gene sequences.

46. The system as recited in claim 36, wherein the program means on said computer uses a local alignment search tool to align the selected gene sequence to each extracted gene sequence.

47. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by amino acid sequences.

48. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by nucleic acid sequences.

49. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by genomic DNA.

50. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by open reading frames.

51. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by introns.

52. The system as recited in claim 36, wherein the selected gene sequence is aligned to each extracted gene sequence by introns and exons.

53. The system as recited in claim 36, wherein the one or more phenotypic criteria excludes genes encoded by mammals.

54. The system as recited in claim 36, wherein the one or more phenotypic criteria excludes genes encoded by zebra fish or frogs.

55. The system as recited in claim 36, wherein the one or more phenotypic criteria excludes genes encoded by invertebrates.

56. The system as recited in claim 36, wherein said system may be used for high specificity primer selection.

57. The system as recited in claim 36, wherein said system may be used for high specificity positioning of gene structures.

58. The system as recited in claim 36, wherein said system may be used for high throughput database conversion.

59. The system as recited in claim 36, wherein said system may be used for high throughput positioning of motifs.

60. A computer program embodied on a computer-readable medium for targeting gene sequences having one or more phenotypic characteristics, said computer program comprising:

first selecting means for selecting one or more phenotypic characteristics of said gene sequences;

second selecting means for selecting a gene sequence that is known to have said one or more of said selected phenotypic characteristics;

third selecting means for selecting at least one database containing cataloged gene sequences therein;

extracting means for extracting from said at least one database a plurality of cataloged gene sequences containing a portion of the said given gene sequence;

aligning means for aligning said given gene sequence to respective ones of said cataloged gene sequence;

prioritizing means for prioritizing the respective ones of the extracted gene sequences based on the alignment of the given gene sequence;

fourth selecting means for selecting at least one of the prioritized gene sequences based on one or more phenotypic criteria; and

designing means for designing one or more degenerate primers to target said at least one selected gene sequence.

61. The computer program as recited in claim 60, further comprising a code segment for filtering the prioritized gene sequences.

62. The computer program as recited in claim 61, wherein the code segment for filtering the prioritized gene sequences removes vertebrate sequences but not invertebrate derived sequences.

63. The computer program as recited in claim 60, further comprising a code segment for cloning genetic material using the one or more degenerate primers.

64. The computer program as recited in claim 60, wherein the one or more databases are selected from cataloged gene sequences for humans, rats, mice, zebra fish, frogs, Drosophila, nematode, C. elegans, mosquito and bacteria.

65. The computer program as recited in claim 60, wherein the phenotypic characteristics include insect mid-gut epithelial cell encoded proteins.

66. The computer program as recited in claim 60, wherein the one or more degenerate primers are nested.

67. The computer program as recited in claim 60, wherein the one or more degenerate primers is used to clone target molecules.

68. The computer program as recited in claim 60, wherein the one or more degenerate primers is used to clone biopesticide encoding genes.

69. The computer program as recited in claim 60, wherein the one or more degenerate primers is used to clone therapeutic encoding genes.

70. The computer program as recited in claim 60, wherein the code segment for prioritizing the extracted gene sequences based on alignment of the selected gene is accomplished by using a statistical analysis of the alignment.

71. The computer program as recited in claim 60, wherein the code segment for prioritizing the extracted gene sequences based on alignment of the selected gene is accomplished by using a local alignment search tool.

72. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by amino acid sequences.

73. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by nucleic acid sequences.

74. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by genomic DNA.

75. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by open reading frames.

76. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by introns.

77. The computer program as recited in claim 60, wherein the selected gene sequence is aligned to each extracted gene sequence by introns and exons.

78. The computer program as recited in claim 60, wherein the one or more phenotypic criteria excludes genes encoded by mammals.

79. The computer program as recited in claim 60, wherein the one or more phenotypic criteria excludes genes encoded by zebra fish or frogs.

80. The computer program as recited in claim 60, wherein the one or more phenotypic criteria excludes genes encoded by invertebrates.

81. An article of manufacture comprising a computer usable medium having computer readable program code means embodied therein for targeting gene sequences, the computer readable program code means in said article of manufacture comprising:

computer readable code means for selecting one or more phenotypic characteristics;

computer readable code means for selecting a gene sequence that is known to have the selected phenotypic characteristics;

computer readable code means for selecting one or more databases containing cataloged gene sequences;

computer readable code means for comparing the selected gene sequence to the cataloged gene sequences;

computer readable code means for extracting any cataloged gene sequences that contain a portion of the selected gene sequence;

computer readable code means for aligning the selected gene sequence to each portion of the extracted gene sequence;

computer readable code means for prioritizing the extracted gene sequences based on the alignment of the selected gene sequence;

computer readable code means for selecting at least one of the prioritized gene sequences based on one or more phenotypic criteria; and

computer readable code means for designing one or more degenerate primers to target the selected-prioritized gene sequences.

82. The article of manufacture of claim 81, wherein said article of manufacture is stored on a medium selected from a group consisting of:

a server, a hard drive, a CD-ROM and a diskette.

**ABSTRACT OF THE DISCLOSURE**

The present invention provides a system, method and apparatus for targeting gene sequences having one or more phenotypic characteristics using a computer. One or more phenotypic characteristics are selected. A gene sequence is then selected that is known to have the selected phenotypic characteristics. In addition, one or more databases containing cataloged gene sequences are selected. The selected gene sequence is compared to the cataloged gene sequences, and any cataloged gene sequences that contain a portion of the selected gene sequence are extracted. The selected gene sequence is aligned to each portion of the extracted gene sequence and the extracted gene sequences are prioritized based on the alignment of the selected gene sequence. At least one of the prioritized gene sequences is selected based on one or more phenotypic criteria. Finally, one or more degenerate primers are designed to target the selected-prioritized gene sequences.

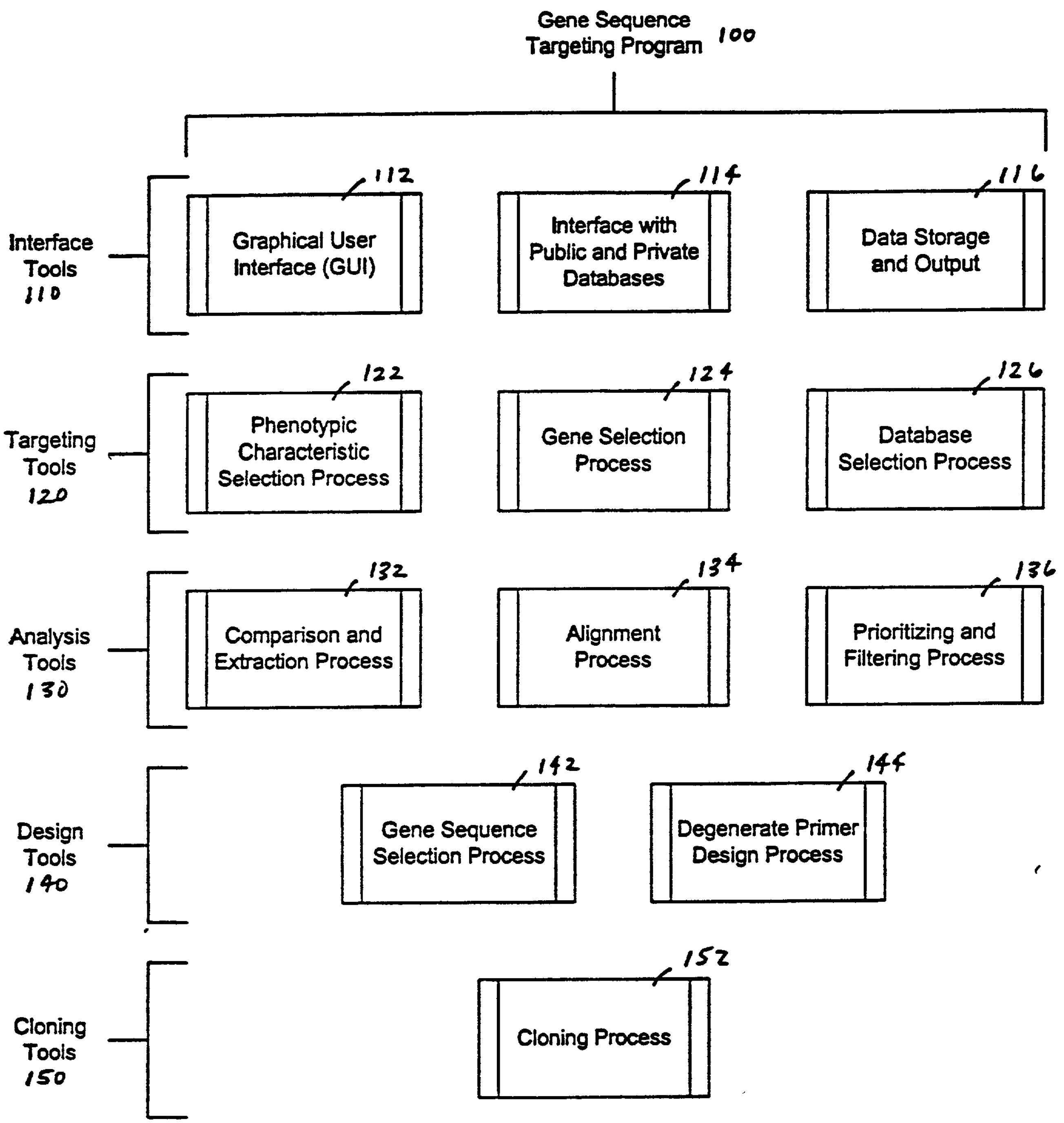

Gene Sequence Targeting Program 100
Interface Tools 110
Graphical User Interface (GUI) 112
Interface with Public and Private Databases 114
Data Storage and Output 116
Targeting Tools 120
Phenotypic Characteristic Selection Process 122
Gene Selection Process 124
Database Selection Process 126
Analysis Tools 130
Comparison and Extraction Process 132
Alignment Process 134
Prioritizing and Filtering Process 136
Design Tools 140
Gene Sequence Selection Process 142
Degenerate Primer Design Process 144
Cloning Tools 150
Cloning Process 152

FIG. 1

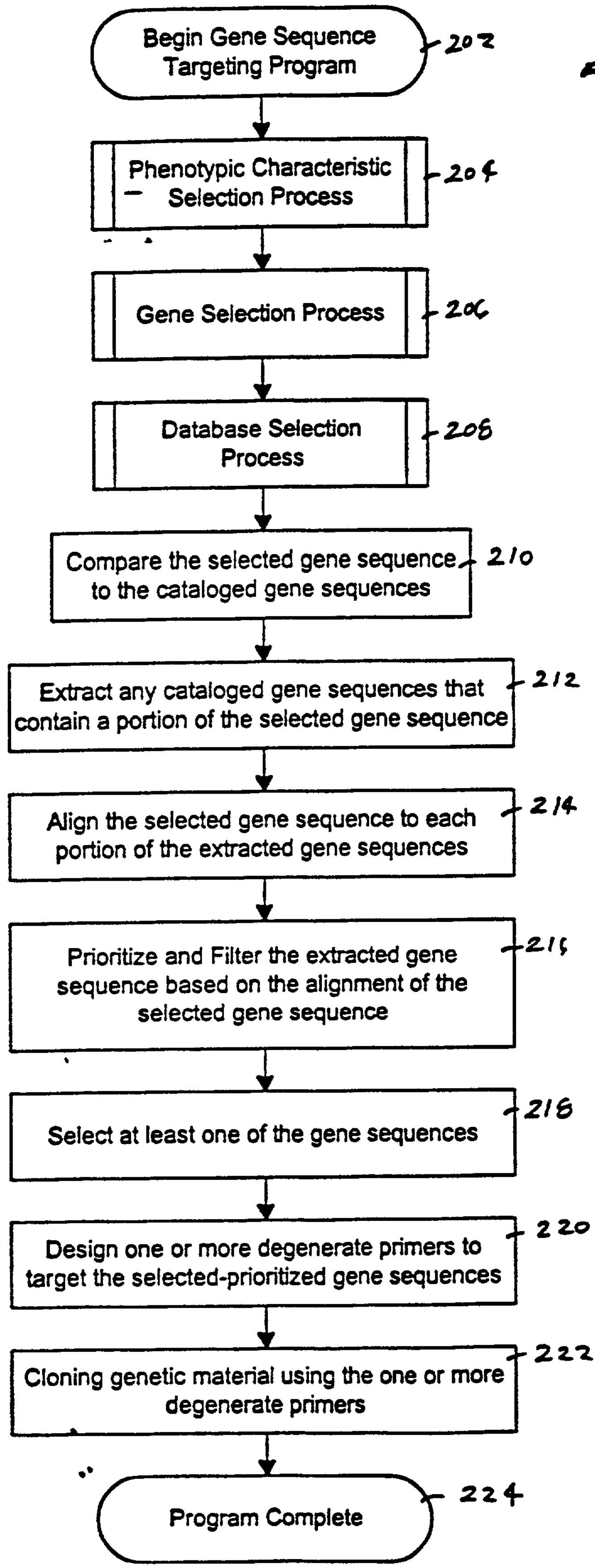

200
Begin Gene Sequence Targeting Program
202
Phenotypic Characteristic Selection Process
204
Gene Selection Process
206
Database Selection Process
208
Compare the selected gene sequence to the cataloged gene sequences
210
Extract any cataloged gene sequences that contain a portion of the selected gene sequence
212
Align the selected gene sequence to each portion of the extracted gene sequences
214
Prioritize and Filter the extracted gene sequence based on the alignment of the selected gene sequence
216
Select at least one of the gene sequences
218
Design one or more degenerate primers to target the selected-prioritized gene sequences
220
Cloning genetic material using the one or more degenerate primers
222
Program Complete
224

FIG. 2

Fig. 2

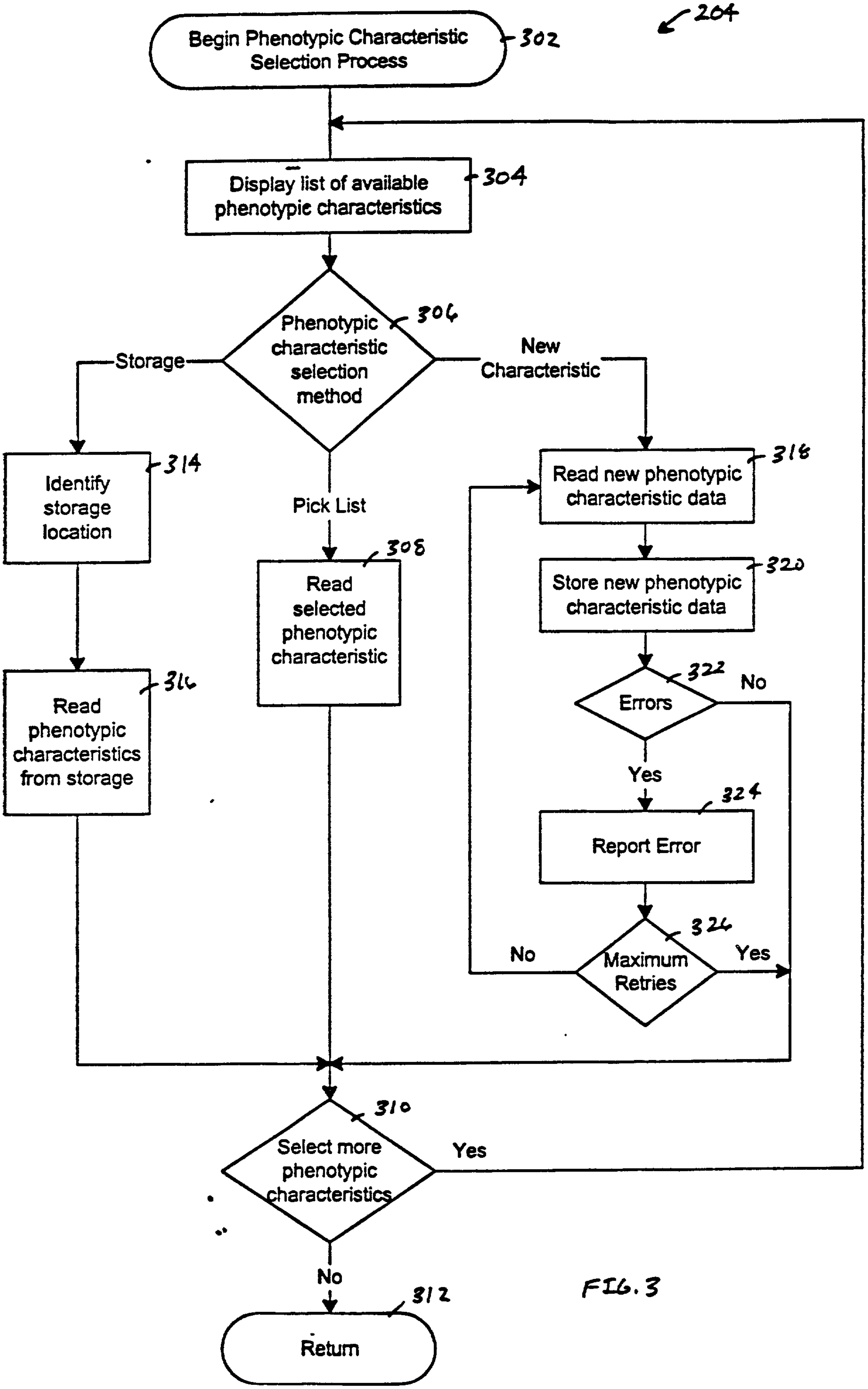
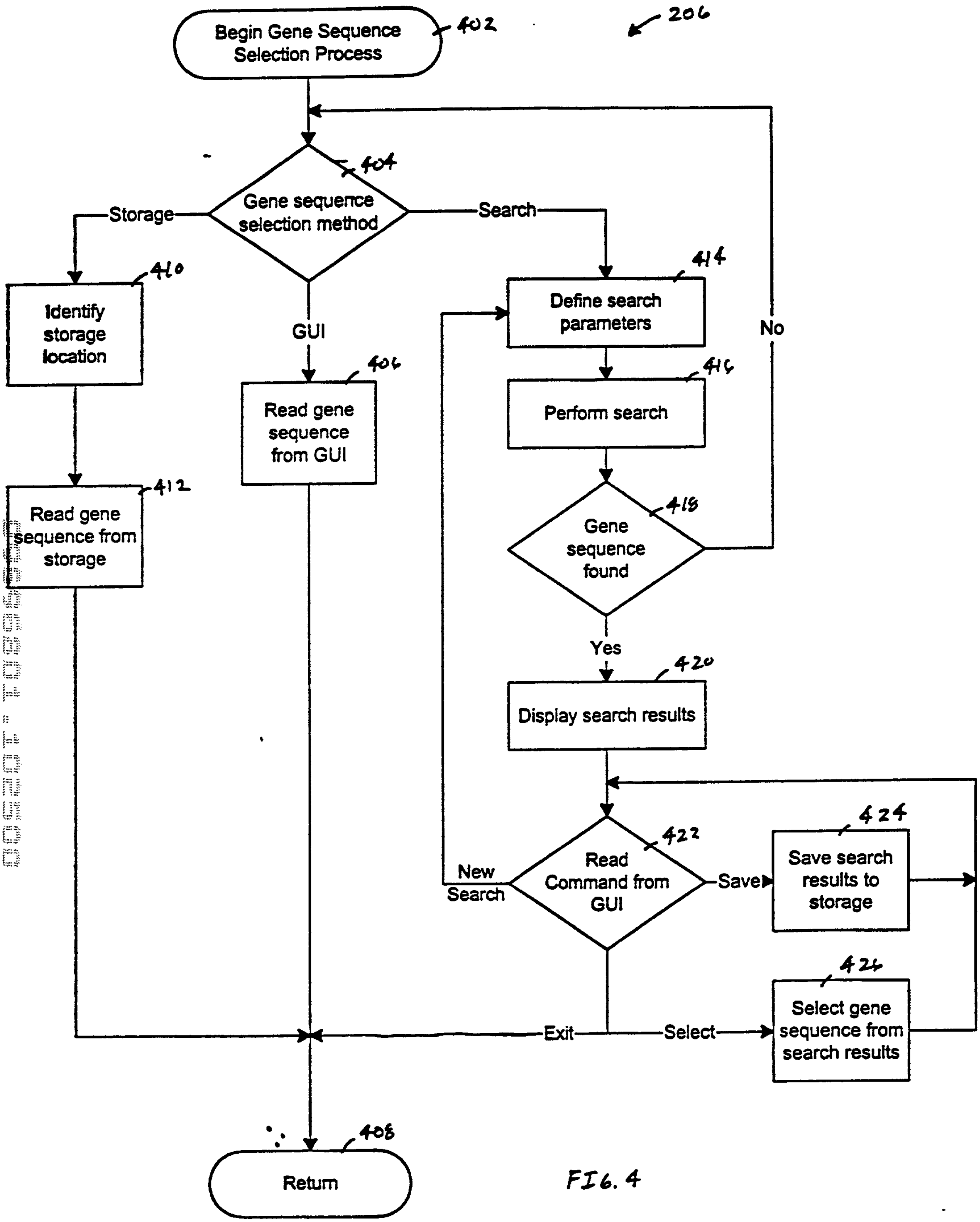

204
Begin Phenotypic Characteristic Selection Process
302
Display list of available phenotypic characteristics
304
Phenotypic characteristic selection method
306
Storage
New Characteristic
Pick List
Identify storage location
314
Read phenotypic characteristics from storage
316
Read selected phenotypic characteristic
308
Read new phenotypic characteristic data
318
Store new phenotypic characteristic data
320
Errors
322
No
Yes
Report Error
324
Maximum Retries
326
No
Yes
Select more phenotypic characteristics
310
Yes
No
Return
312

FIG. 3



206
Begin Gene Sequence Selection Process
402
Gene sequence selection method
404
Storage
Search
GUI
Identify storage location
410
Read gene sequence from GUI
406
Define search parameters
414
Perform search
416
No
Read gene sequence from storage
412
Gene sequence found
418
Yes
Display search results
420
Read Command from GUI
422
New Search
Save
Save search results to storage
424
Select gene sequence from search results
426
Exit
Select
Return
408


FIG. 4

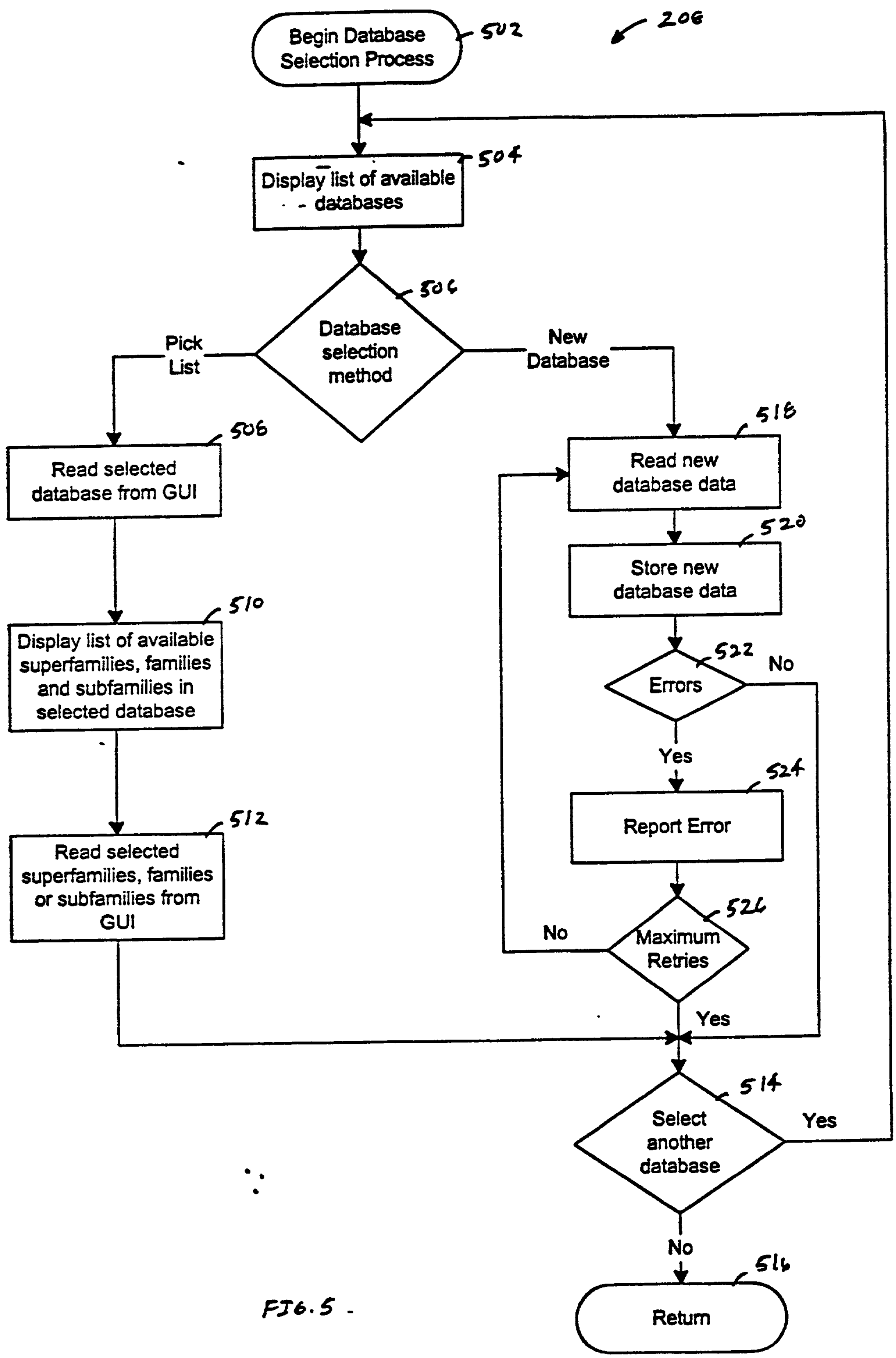

Begin Database Selection Process
502
206
Display list of available databases
504
Database selection method
506
Pick List
New Database
Read selected database from GUI
508
Read new database data
518
Store new database data
520
Display list of available superfamilies, families and subfamilies in selected database
510
Errors
522
No
Yes
Report Error
524
Read selected superfamilies, families or subfamilies from GUI
512
Maximum Retries
526
No
Yes
Select another database
514
Yes
No
Return
516

FIG. 5

# System Network Overview

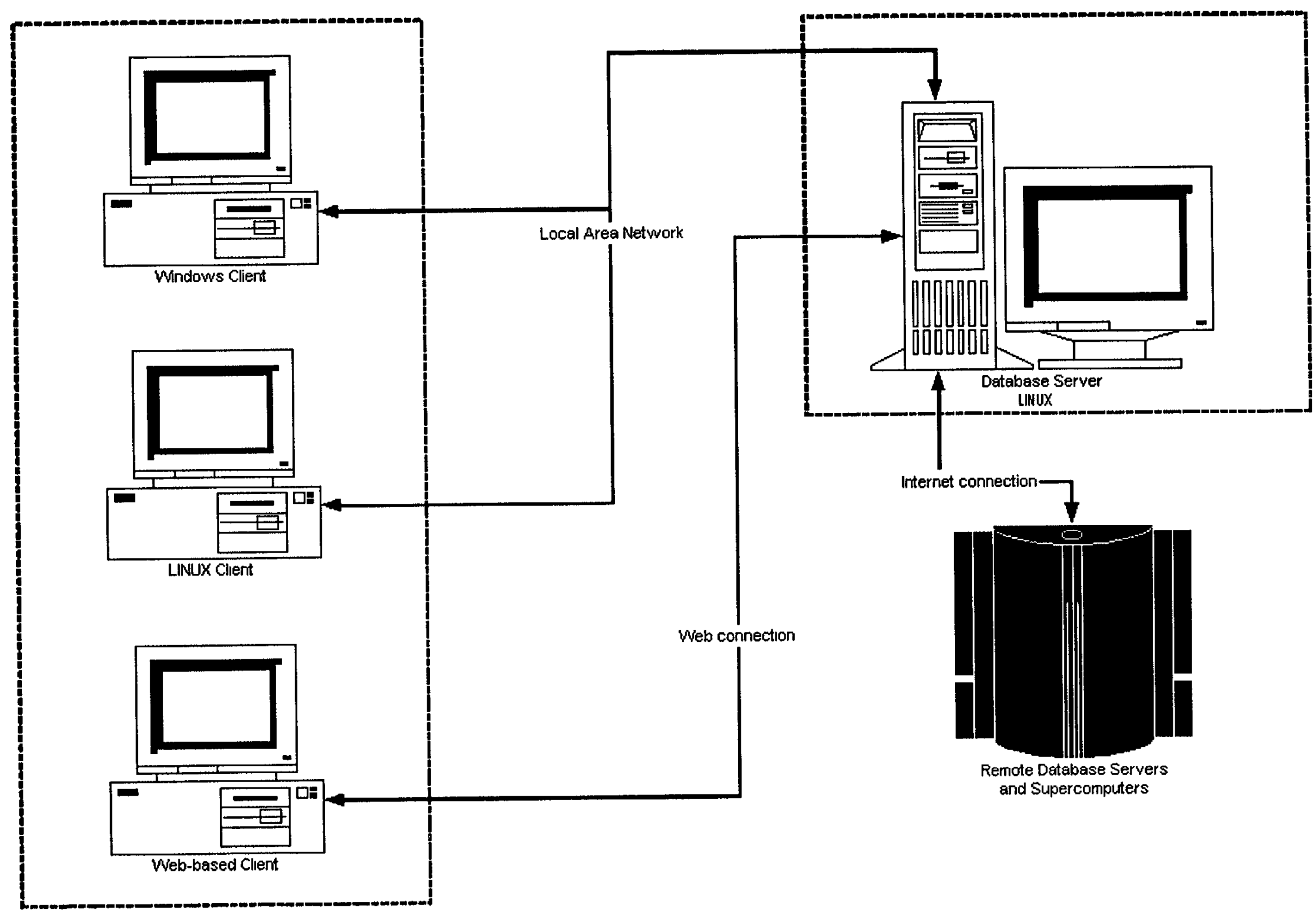



FIGURE 6

# Database Management

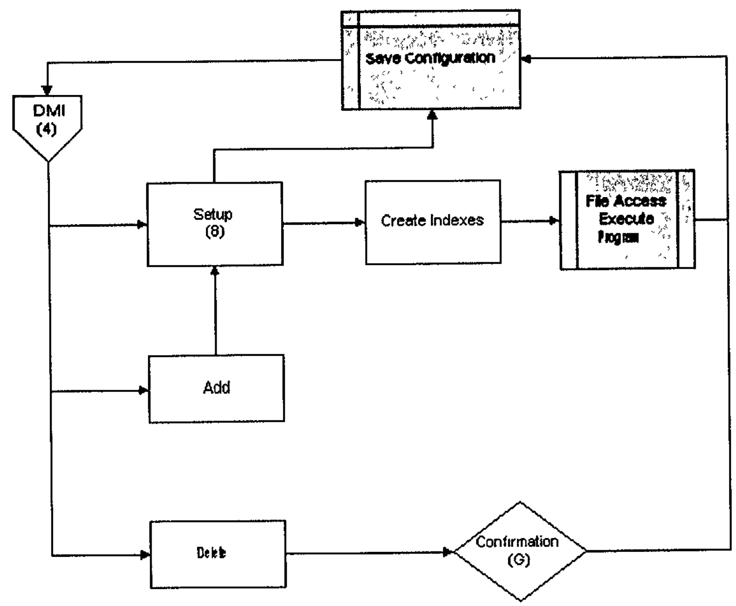



FIGURE 8

# Workspace Management

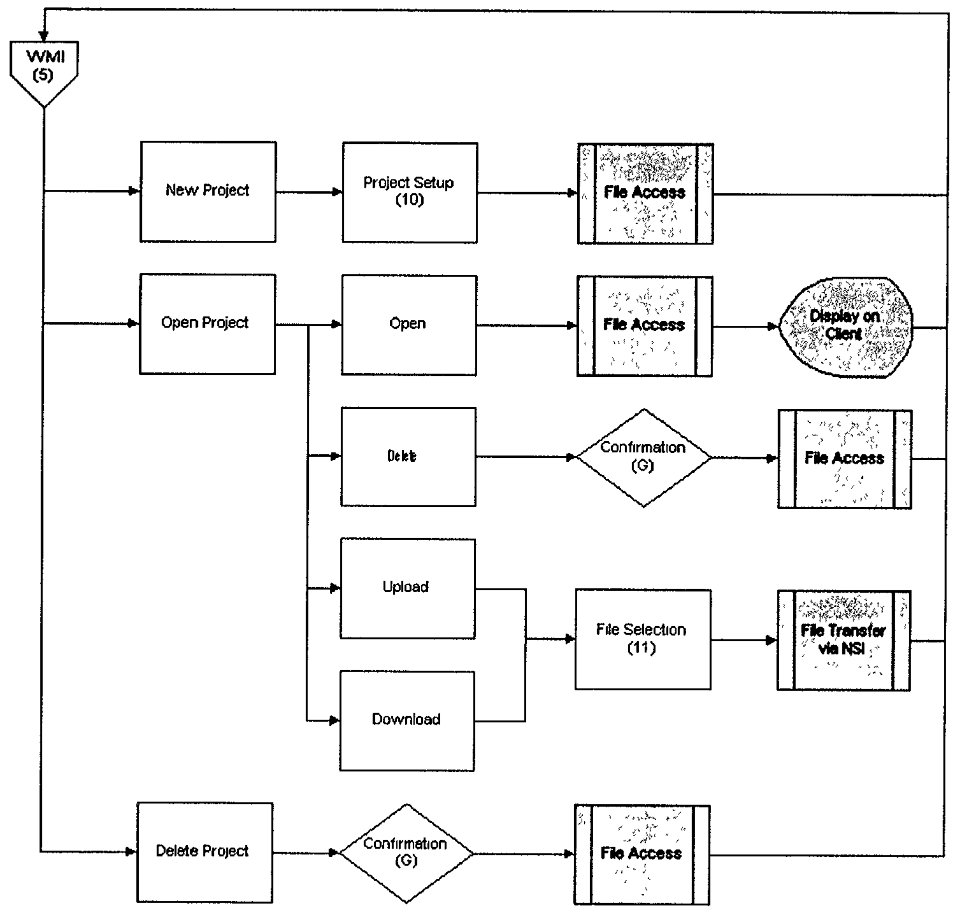



FIGURE 9

# Search Analysis Tools

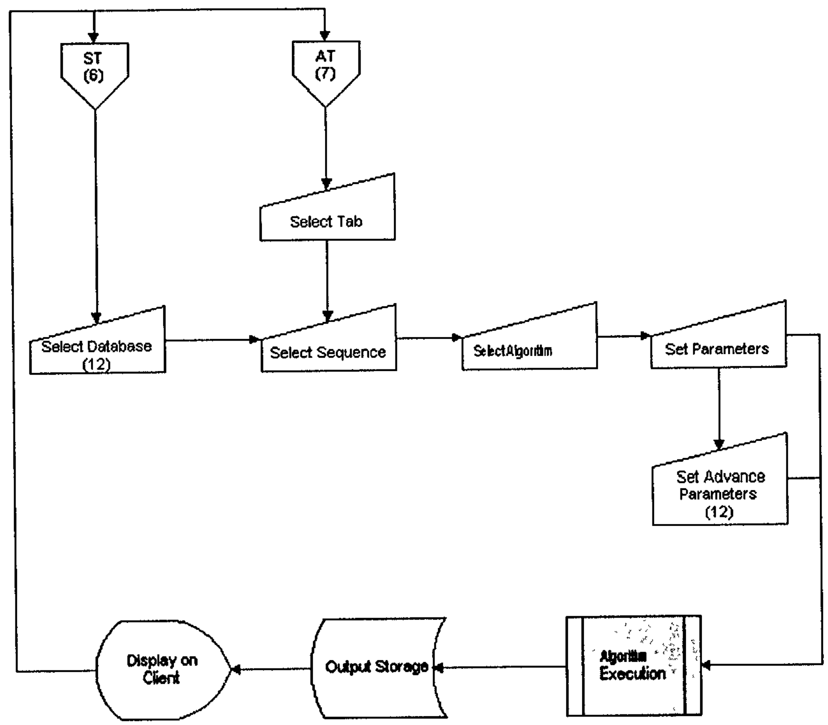



FIGURE 10

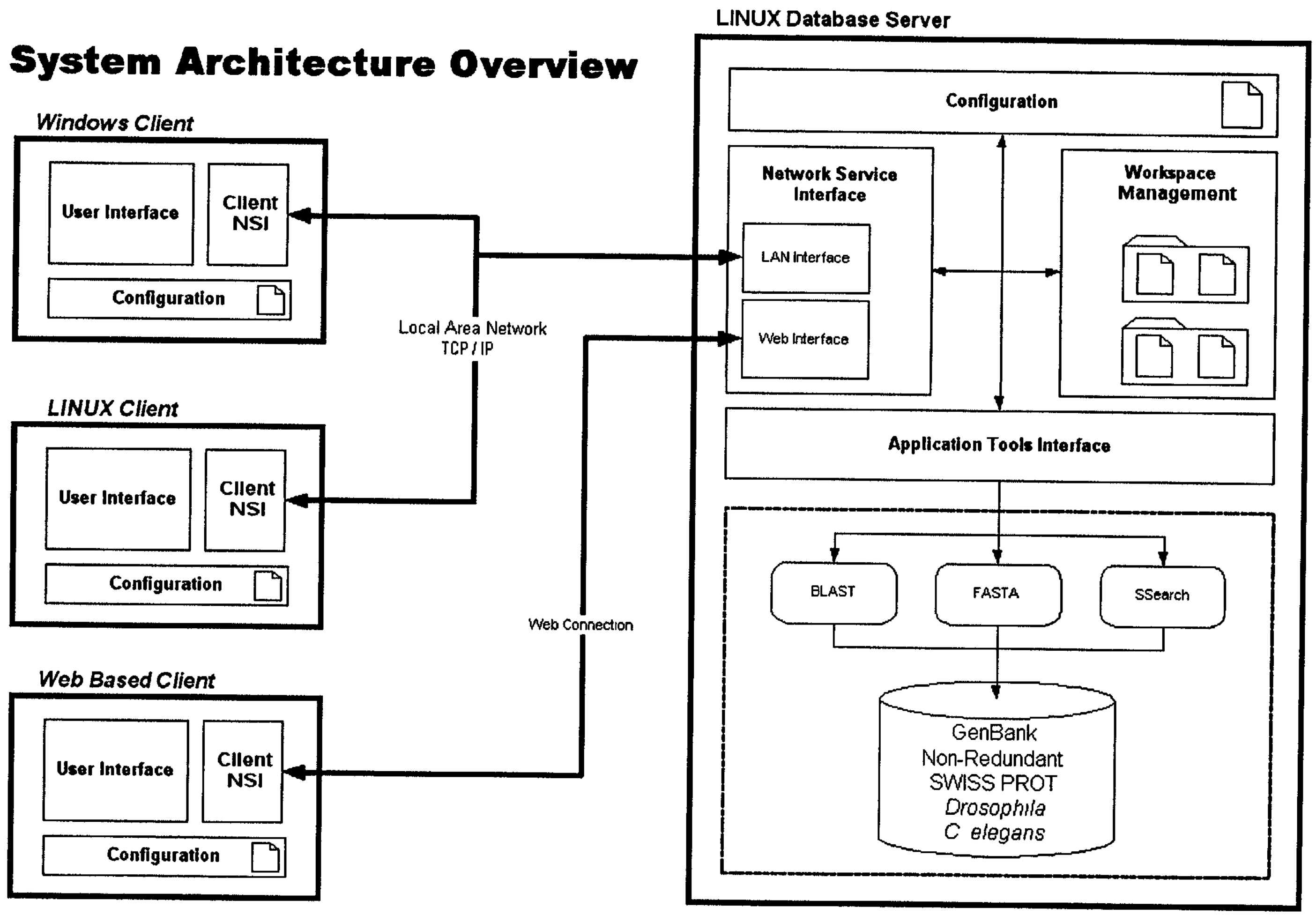

System Architecture Overview
Windows Client
User Interface
Client NSI
Configuration
LINUX Client
User Interface
Client NSI
Configuration
Web Based Client
User Interface
Client NSI
Configuration
Local Area Network
TCP / IP
Web Connection
LINUX Database Server
Configuration
Network Service Interface
LAN Interface
Web Interface
Workspace Management
Application Tools Interface
BLAST
FASTA
SSearch
GenBank
Non-Redundant
SWISS PROT
Drosophila
C elegans

FIGURE 11

# Process Flow

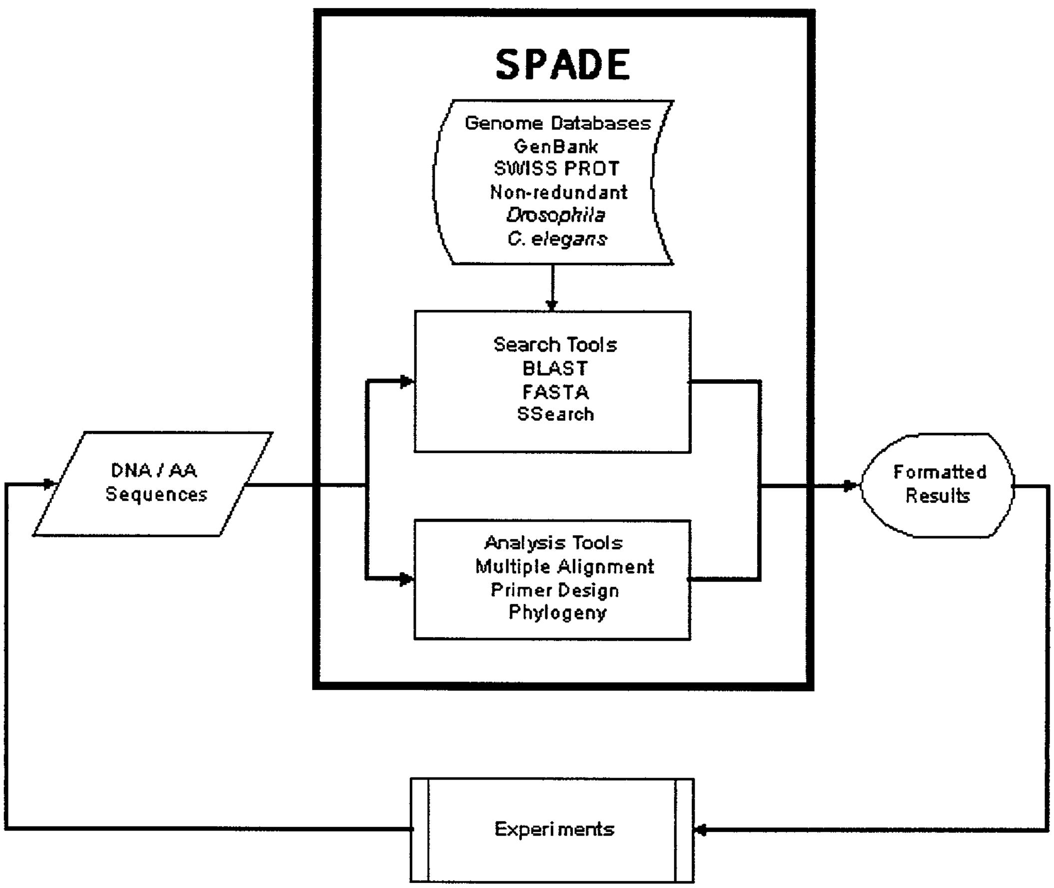



**FIGURE 12**

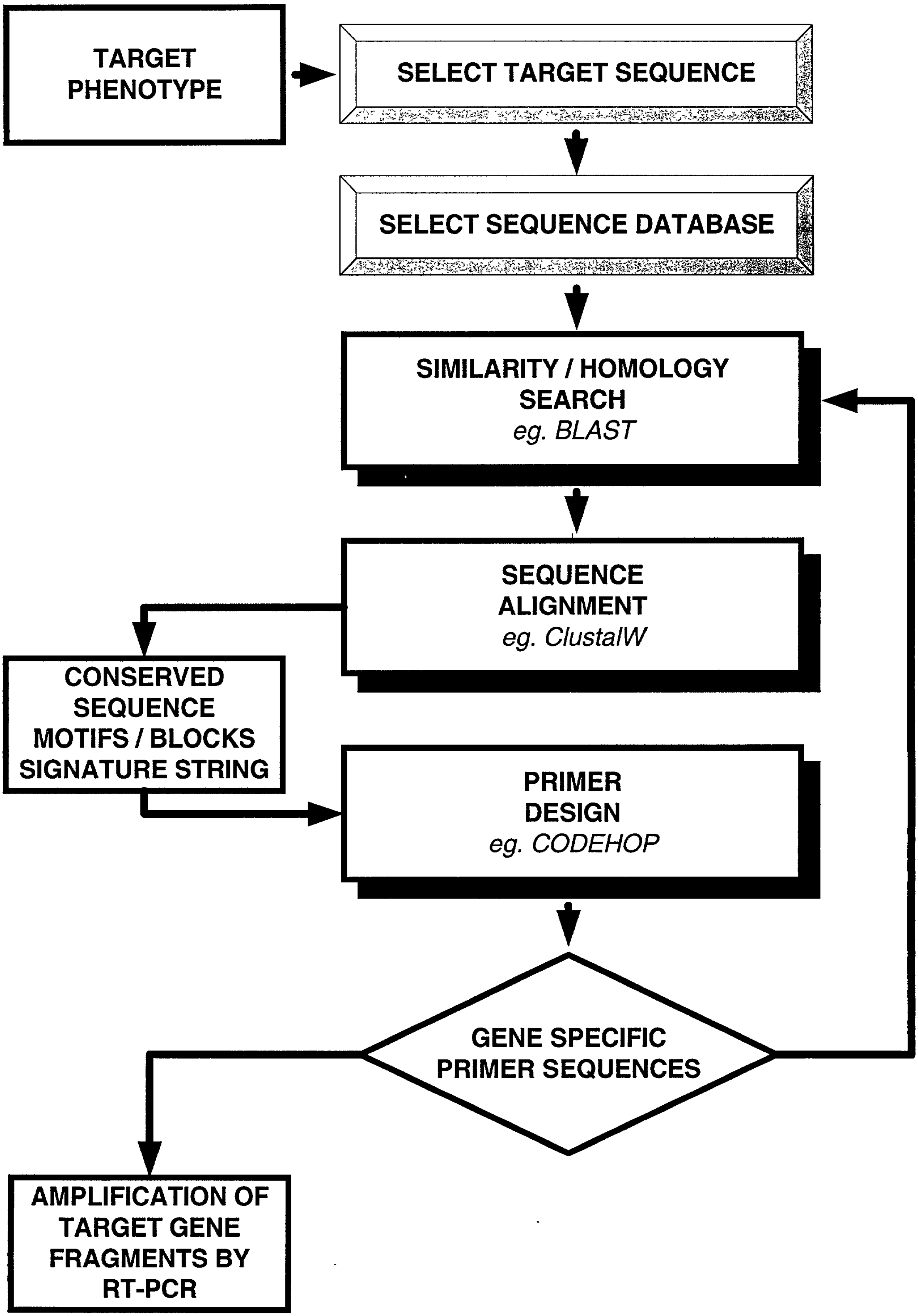

TARGET PHENOTYPE
SELECT TARGET SEQUENCE
SELECT SEQUENCE DATABASE
SIMILARITY / HOMOLOGY SEARCH
eg. BLAST
SEQUENCE ALIGNMENT
eg. ClustalW
CONSERVED SEQUENCE MOTIFS / BLOCKS SIGNATURE STRING
PRIMER DESIGN
eg. CODEHOP
GENE SPECIFIC PRIMER SEQUENCES
AMPLIFICATION OF TARGET GENE FRAGMENTS BY RT-PCR

FIGURE 13

```
AGGGAGGATTCGATGCAATAATGCAAGTCATGACTTGTGAGAAAGAAATAGGGTGGAGACCTGGCTCAAG 70
K  G  G  F  D  A  I  M  Q  V  M  T  C  E  K  E  I  G  W  R  P  G  S  R
GCGTATAATTGTTCTGTGCACCGATTCCCCATATCACAGCGCTGGTGACGGCAAAATGATAGGCATTATC 140
 R  I  I  V  L  C  T  D  S  P  Y  H  S  A  G  D  G  K  M  I  G  I  I
AAACCCAACGACATGTTATGCCACTTAAAGGGACAAAAATATGAAGCAGAAATGGCCCAAGATTATCCAT 210
 K  P  N  D  M  L  C  H  L  K  G  Q  K  Y  E  A  E  M  A  Q  D  Y  P
CTGTGAGTAAAATAAATAAAGTAGCAAAGCAAGGAAAATTCGGTATCATATTCGCTGCTTTGGCTGAGGT 280
S  V  S  K  I  N  K  V  A  K  Q  G  K  F  G  I  I  F  A  A  L  A  E  V
CCGTGATGTTTATACCTTGTTAGCGGAACAAATAGTCGGAGCTGAGTACGCCGAACTGAAGAAACAGAAG 350
 R  D  V  Y  T  L  L  A  E  Q  I  V  G  A  E  Y  A  E  L  K  K  Q  K
TCAAATATTGTAGAGATCATTATAAAAGCGTACCAACGCAGCGTTCGAAGTATCAAATTGGATTATGACA 420
 S  N  I  V  E  I  I  I  K  A  Y  Q  R  S  V  R  S  I  K  L  D  Y  D
TACCTTCATTCGTTAGACTGAAACTTAATCAAAGTTGTGACGGGACACCAATTAATTGTGCCAGCACCTA 490
I  P  S  F  V  R  L  K  L  N  Q  S  C  D  G  T  P  I  N  C  A  S  T  Y
TGAAAATCCAGTGGTTACAATTCCGGCTATTCTAGAGGTTAAAGAATGTCCTAAAGAAAATAAAACACAT 560
 E  N  P  V  V  T  I  P  A  I  L  E  V  K  E  C  P  K  E  N  K  T  H
GAGCTTGTTATTAACCCTGTGTCTTTAAATGACAAATTAATAATTAAATTGGAAGTCATCTGTAAATGTG 630
 E  L  V  I  N  P  V  S  L  N  D  K  L  I  I  K  L  E  V  I  C  K  C
ACTGTGAAGTCAAAAGTGATATAAGTTCAAGATGTAATAATGCAGGATATATACAGTGTGGTATCTGCAA 700
D  C  E  V  K  S  D  I  S  S  R  C  N  N  A  G  Y  I  Q  C  G  I  C  K
GTGTCTCGATTCAAGTTATGGCGACGAATGTCAGTGCAGCGTTACATCTTCGGGGGTGGCTAATAAGGAG 770
 C  L  D  S  S  Y  G  D  E  C  Q  C  S  V  T  S  S  G  V  A  N  K  E
AAAGATGACGCCAAATGCCGTAAGGATCTAAATGACATAGTACTGTGTAGTGGGAAAGGCGTATGTATGT 840
 K  D  D  A  K  C  R  K  D  L  N  D  I  V  L  C  S  G  K  G  V  C  M
GTGGTAAATGTACTTGTAACCCTGATCGTTCAGGAAAATATTGCGAATTTGACGATAAGGCATGCGATAA 910
C  G  K  C  T  C  N  P  D  R  S  G  K  Y  C  E  F  D  D  K  A  C  D  N
TCTTTGCTCAAACCATGGGATTTGTACCTTAGGCTCATGCCAGTGCGATAGCGGTTGGTCTGGAAATGAT 980
 L  C  S  N  H  G  I  C  T  L  G  S  C  Q  C  D  S  G  W  S  G  N  D
TGCGGTTGTCCAACTAGTAACACAGACTGCTACGCTCAATACTCTGAGGAGGTTTGTTCTGGCAACGGTG 1050
 C  G  C  P  T  S  N  T  D  C  Y  A  Q  Y  S  E  E  V  C  S  G  N  G
AATGTGTATGCGGAAATGCCAATGTGCGAAGGTTAAAGGAAAAAACGAAACGTACGCAGGAGTATTTTGT 1120
E  C  V  C  G  N  A  N  V  R  R  L  K  E  K  T  K  R  T  Q  E  Y  F  V
GACACATGCAATGACTGCCAATCAAAATATTGTAAAGCCCTCGAACCAAATGTAGAATGTAACTACATAC 1190
 T  H  A  M  T  A  N  Q  N  I  V  K  P  S  N  Q  M  .  N  V  T  T  Y
AAGGTCTAGAAACTTGTGATAAGATTTACAACAATACAGAAAACAATGTTGTTATAAAAATGGTCAACAA 1260
 K  V  .  K  L  V  I  R  F  T  T  I  Q  K  T  M  L  L  .  K  W  S  T
AACAGAAATTAATTCGCCTAAATGGAGTGGCGCTACTTGGTGCAAAAAAGTAATAGAGGACGGCAGTTTT 1330
K  Q  K  L  I  R  L  N  G  V  A  L  L  G  A  K  K  .  .  R  T  A  V  L
ATAATATTCAGATATTATCATAACGCAACGACGCACGGGTTGCATATAATCATTCAAACGGAACCAGAGG 1400
 .  Y  S  D  I  I  I  T  Q  R  R  T  G  C  I  .  S  F  K  R  N  Q  R
CACCTCCAATAGGAAATAAGTGGATTGCCCTCATCAGTTGCATAGTGGCTGTAGTACTCATTGGCTTGTT 1470
 H  L  Q  .  E  I  S  G  L  P  S  S  V  A  .  W  L  .  Y  S  L  A  C
GACGTTGATTGCGTGGAAGATCCTCGTAGACTTGCACGATAAAAGGGAATATGCCAAGTTGA 1532
.  R  .  L  R  G  R  S  S  .  T  C  T  I  K  G  N  M  P  S  .
```

FIGURE 14

**RULES 63 AND 67 (37 C.F.R. 1.63 and 1.67)**
**DECLARATION AND POWER OF ATTORNEY**

FOR UTILITY/DESIGN/CIP/PCT NATIONAL APPLICATIONS

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name; and

I believe that I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled: **GENE MINING SYSTEM AND METHOD,** the specification of which: (mark only one)

_x_ (a) is attached hereto.
___ (b) was filed on __________ as Application Serial No. __________ and was amended on __________________ (if applicable)
___ (c) was filed as PCT International Application No. PCT/_________ on ______ _____ and was amended on ______________ (if applicable).
___ (d) was filed on ______________ as Application Serial No. _________and was issued a Notice of Allowance on ________________.
___ (e) was filed on _________ and bearing attorney docket number _______.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims as amended by any amendment referred to above or as allowed as indicated above.

I acknowledge the duty to disclose all information known to me to be material to the patentability of this application as defined in 37 CFR § 1.56. If this is a continuation-in-part (CIP) application, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose to the Office all information known to me to be material to patentability of the application as defined in 37 CFR § 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this CIP application.

I hereby claim foreign priority benefits under 35 U.S.C. § 119/365 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate filed by me or my assignee disclosing the subject matter claimed in this

application and having a filing date (1) before that of the application on which my priority is claimed or, (2) if no priority is claimed, before the filing date of this application:

PRIOR FOREIGN PATENTS

| Number | Country | Month/Day/Year Filed | Date first laid-open or Published | Date patented or Granted | Priority Claimed Yes | No |
|---|---|---|---|---|---|---|
| _____ | _____ | _____ | _____ | _____ | _____ | _____ |

I hereby claim the benefit under 35 U.S.C. § 120/365 of any United States application(s) listed below and PCT international applications listed above or below:

PRIOR U.S. OR PCT APPLICATIONS

| Application No. (series code/serial no.) | Month/Day/Year Filed | Status(pending, abandoned, patented) |
|---|---|---|
| 60/161,527 | October 26, 1999 pending | |
| 60/161,571 | October 26, 1999 pending | |

I hereby appoint:

Timothy G. Ackermann, Reg. No. 44,493
Benjamin J. Bai, Reg. No. 43,481
Michael J. Blankstein, Reg No. 37,097
Mary Jo Boldingh, Reg. No. 34,713
Margaret A. Boulware, Reg. No. 28,708
Arthur J. Brady, Reg. No. 42,356
Matthew O. Brady, Reg. No. 44,554
Daniel J. Burnham, Reg. No. 39,618
Thomas L. Cantrell, Reg. No. 20,849
Ronald B. Coolley, Reg. No. 27,187
Thomas L. Crisman, Reg. No. 24,846
Stuart D. Dwork, Reg. No. 31,103
William F. Esser, Reg. No. 38,053
Roger J. French, Reg. No. 27,786
Janet M. Garetto, Reg. No. 42,568
John C. Gatz, Reg. No. 41,774
Russell J. Genet, Reg. No. 42,571
Gerald H. Glanzman, Reg. No. 25,035
J. Kevin Gray, Reg. No. 37,141
Steven R. Greenfield, Reg. No. 38,166
Joshua A. Griswold, Reg. No. 46,310
J. Pat Heptig, Reg. No. 40,643
Sharon A. Israel, Reg. No. 41,867
John R. Kirk Jr., Reg. No. 24,477
Paul R. Kitch, Reg. No. 38,206
Timothy M. Kowalski, Reg. No. 44,192
James F. Lea Iii, Reg. No. 41,143
Hsin-wei Luang, Reg. No. 44,213
Robert W. Mason, Reg. No. 42,848
Roger L. Maxwell, Reg. No. 31,855
Robert A. Mcfall, Reg. No. 28,968
Steven T. Mcdonald, Reg. No. 45,999
Lisa H. Meyerhoff, Reg. No. 36,869
Stanley R. Moore, Reg. No. 26,958
Richard J. Moura, Reg. No. 34,883
Mark V. Muller, Reg. No. 37,509
P. Weston Musselman Jr. Reg No. 31,644
Daniel G. Nguyen, Reg. No. 42,933
Spencer C. Patterson, Reg. No. 43,849
Russell N. Rippamonti, Reg. No. 39,521
Ross T. Robinson, Reg. No. 47,031
Stephen G. Rudisill,, Reg. No. 20,087
Holly L. Rudnick, Reg. No. 43,065
J.I. Jennie Salazar, Reg. No. 45,065
Keith W. Saunders, Reg. No. 41,462
Jerry R. Selinger, Reg. No. 26,582
Gary B. Solomon, Reg. No. 44,347
Steve Z. Szczepanski, Reg. No. 27,957
Andre M. Szuwalski, Reg. No. 35,701
Alan R. Thiele, Reg. No. 30,694
Tamsen Valoir, Reg. No. 41,417
Raymond Van Dyke, Reg. No. 34,746
Brian D. Walker, Reg. No. 37,751
Gerald T. Welch, Reg. No. 30,332
Harold N. Wells, Reg. No. 26,044
William D. Wiese, Reg. No. 45,217

all of the firm of **JENKENS & GILCHRIST, P.C.**, 3200 Fountain Place, 1445 Ross Avenue, Dallas, Texas 75202-2799, as my attorneys and/or agents, with full power of substitution and revocation, to prosecute this application, provisionals thereof, continuations, continuations-in-part, divisionals, appeals, reissues, substitutions, and extensions thereof and to transact all business in the United States Patent and Trademark Office connected therewith, to appoint any individuals under an associate power of attorney and to file and prosecute any international patent application filed thereon before any international authorities, and I hereby authorize them to act and rely on instructions from and communicate directly with

the person/assignee/attorney/firm/organization who/which first sent this case to them and by whom/which I hereby declare that I have consented after full disclosure to be represented unless/until I instruct them in writing to the contrary.

Please address all correspondence and direct all telephone calls to:

Raymond Van Dyke, Esq.
Jenkens & Gilchrist, P.C.
3200 Fountain Place
1445 Ross Avenue
Dallas, Texas 75202-2799
214/855-4708    214/855-4300 (fax)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

**NAMED INVENTOR(S)**

<table>
<tr><td rowspan="3">1</td><td>Lee A. Bulla, Jr.<br>Full Name</td><td>Inventor's Signature</td><td>25Oct00<br>Date</td></tr>
<tr><td colspan="2">6455 U.S. Hwy. 377<br>Tioga, Texas 76271<br>Residence (city, state, country)</td><td>US<br>Citizenship</td></tr>
<tr><td colspan="3">6455 U.S. Hwy. 377<br>Tioga, Texas 76271<br>Post Office Address (include zip code)</td></tr>
<tr><td rowspan="3">2</td><td>Mehmet Candas<br>Full Name</td><td>Inventor's Signature</td><td>10/25/2000<br>Date</td></tr>
<tr><td colspan="2">13859 Peyton Dr. 1<br>Dallas, Texas 75240<br>Residence (city, state, country)</td><td>US<br>Citizenship</td></tr>
<tr><td colspan="3">13859 Peyton Dr. 1<br>Dallas, Texas 75240<br>Post Office Address (include zip code)</td></tr>
</table>

United States Patent & Trademark Office

Office of Initial Patent Examination -- Scanning Division

SCANNED, # ____

Application deficiencies were found during scanning:

☐ Page(s) 1 of Drawings were not present for scanning.
(Document title)

☐ Page(s) ________ of ____________________ were not present for scanning.
(Document title)

☐ Scanned copy is best available.